THE UNIVERSITY
of EDINBURGH

PRC1 organizes 3D chromatin architecture in mouse ES cells

THE UNIVERSITY *of* EDINBURGH

Ilya M. Flyamer

Thesis presented for the degree of Doctor of Philosophy

The University of Edinburgh

2019

Declaration

I declare that this thesis has been composed by me, and that all the work is my own unless otherwise stated.

This work has not been submitted for any other degree.

Ilya Flyamer

September 2019

Acknowledgements

A large number of people contributed to this work to make it happen the way that it did.

First of all, I am grateful to Wendy for guidance, enthusiasm for science and support, including paying for all the costly Hi-C sequencing. I would also like to thank Richard for co-supervising of the 2i project and always being friendly and supportive.

Thank you Katy, for letting me jump onto the 2i ship, sharing in the paper frustration, and for answering my questions about Hox expression in 2i about 50 times!

Rob was my *de facto* supervisor for most of the PhD with vast knowledge of all things polycomb, DNA methylation and sequencing, which were key for my project – thank you for that, and for the great chat! Good luck with your own lab, and hope we will continue collaborating in the future!

I am also grateful to Iain, my original day to day supervisor who taught me FISH, and a fellow C-guy. Hope you are not too annoyed at my constant reminders that Hi-C/capture-Hi-C is waaay better than 5C (but it's true)!

Shelagh deserves a special mention here as the main source of amazing cake together with advice on the Scottish life, language and food. And very well timed coffee breaks and Dishoom trips. And scones, scone advice and scone inspiration. And also thank you for keeping the lab running and all the FISH expertise! While FISH was not included in this thesis, the whole RING1B

project wouldn't have happened without approximately 783 years worth of your FISH experiments.

A lot of other bickmoreans have been a great help during the course of my PhD. Gabi showed me how to do Westerns, Nezha explained the TALe assembly, Charlene helped with immuno-FISH, and I am probably forgetting other important influences right now. And I am sure everyone shared a reagent or two with me at some point!

Johanna and Sergey in Vienna and Moscow, respectively, have been great collaborators and friends, always fun to chat to, about science or not, and share freshly discovered papers.

Pairtoolers, mainly from the Mirny lab, have been a great help with all things Hi-C analysis, in particular Nezar, Anton and Max – you guys are .cool!

Thank you to the lunch group (Nefeli, Lana, Toby throughout my time here, with also Anna and Lara in the beginning, and Thomas and Katerina later on) for a consistent lunchtime at 12:15 and conversations full of amazing facts and hilarious jokes (Toby, looking at you). Also, thank you to the WGH canteen for quenching the hunger with real food and at times of day when there is no other option!

Kathryn, you were a great office neighbour, but sadly now on the dark side and we don't talk as much – but your wedding has been all planned out and went smoothly, so this was probably for the best in the interest of the productivity in this half of the office, although your Russian skills have regressed significantly. At least we still play more or, preferably, less frustrating legacy board games together with Matthew and Lauren when we get a chance.

Thank you to Agniete, Маша, Yavor, Karolina and other EUTS dancers for always being there on the dance floor at milongas and classes alike, helping me improve my tango skills, and dinners and drinks.

While mentioned before as part of the lunch group, I have to mention Nefeli again as a special friend with whom we have shared a lot of adventures already, and I am sure will share much more.

Finally, I want to thank my mum for not only supporting my curiosity before and during my PhD, but also for hard cash help first with my move to Edinburgh, and then again to allow me to stay here.

Abstract

Polycomb is a major epigenetic pathway involved in developmental gene regulation. While currently it is not understood how Polycomb Repressive Complexes (PRCs) silence genes, regions of the genome bound by polycomb are highly compacted, and cluster together in nuclear space. An appealing model is that this structural property of polycomb complexes contributes to gene silencing.

To investigate this, here I apply Hi-C to analyse chromatin organization in 3D in mouse embryonic stem cells (mESCs) lacking RING1B, a core component of PRC1, and in cells with impaired catalytic function of RING1B. Hi-C is a molecular method coupled to high-throughput sequencing that can interrogate the 3D organisation of genomes.

The main approach used to quantify enrichment interactions in Hi-C datasets is creation of so-called pile-ups: averaging of multiple 2D regions of the Hi-C map to visualize the average profile of interactions between features of interest. Pile-ups are a crucial analytical approach, however there was no convenient tool available for the task. Therefore, first I developed my own versatile tool - *coolpup.py*. I show that *Coolpup.py* works for extremely sparse single-cell Hi-C data. Moreover, I have used *coolpup.py* to discover a novel dynamic pattern of polycomb-associated loops during cell cycle progression that is a completely different pattern from CTCF-mediated loops, with most prominent polycomb-mediated interaction enrichment occurring just before and just after mitosis.

Second, using computational analysis including application of *coolpup.py*, I describe properties of polycomb-mediated chromatin structures and the role of

PRC1 in creating them, in mouse embryonic stem (ES) cells. Using a RING1B knock-out line, I showed that the PRC1 complex creates both local compaction of its targets, and distal interaction between them. Using cells with a RING1B I53A point mutation which has greatly impaired E3 ubiquitin ligase activity, my data suggest that there is no direct role of the H2AK119 ubiquitination in formation of 3D chromatin structures. I investigate what determines formation of distal interactions between RING1B binding sites. Surprisingly, distance turned out not to be an important factor, since I detect enriched interactions even at distances as large as 50-100 Mbp. I detect a clear role of canonical PRC1 binding, unlike PRC2 or variant PRC1 complexes in forming these interactions. I show that transcriptional activation of polycomb targets in cells lacking RING1B is not required for loss of interactions between them.

Finally, I have investigated 3D genome re-organization in ES cells grown in conditions of ground state pluripotency in “2i” medium. I have shown that 2i conditions lead to a depletion of chromatin compaction and looping mediated by polycomb. Moreover, I have shown that the PRC1-mediated interactions, while weakened, are still present in 2i grown cells, unlike the inner cell mass where no interactions can be observed.

Overall, I have demonstrated the abundance of PRC1-mediated structures present in the ES cell nucleus, the role of canonical PRC1 complexes in their formation, the plasticity of the 3D genome organization which follow epigenome changes, during both cell cycle progression and upon entering ground state pluripotency.

Lay summary

Activity of genes is tightly regulated during development of an organism. While it is crucial to activate the correct genes in correct cells, it is also very important to silence genes that should not be active, both in the wrong tissue and at a wrong time in development. I am particularly interested in how gene regulation is related to the folding of DNA in 3D in the cell nucleus.

Here I first developed a computational tool I then used to quantify structure of the DNA in genome-wide experimental data. Then I describe how one of the major gene repression systems operates to fold the cell DNA: the silenced genes are found close together in 3D, and they are highly compacted. I confirmed that these structures are generated by a particular group of repressive molecular complexes by using cells lacking one of its key components. This 3D organization of DNA potentially can be one of the mechanisms of gene repression used by this system. Finally, I described the changes in this 3D structures observed in different culture conditions of embryonic stem cells, that model different stages of early development. Surprisingly, I found very high level of plasticity of these structures in this model, indicating that they are not required for proper cellular function at these early stages of development.

Table of Contents

List of Figures

List of Tables

Chapter 1: Introduction

1.1 Chromatin structure and function

In mammalian cells the ~2 metre long genome is packed inside of a nucleus ~10 µm in diameter. The genome has to perform its functions, such as supporting transcription, and at the same time maintain the high level of compaction required to fit inside the nucleus. Below I review what is known about how this is achieved in mammalian cells, and how 3D folding of the genome is thought to be important for the nuclear function.

1.1.1 The nucleosome level of chromatin organization

In the nuclei of eukaryotic cells DNA forms a nucleoprotein complex called chromatin. The main protein components of chromatin are the nucleosomes formed by octamers of histones. The DNA in the nucleus is wrapped around the histone octamers, forming the so-called 10-nm fiber, the lowest level of chromatin organization. Each nucleosome consist of 2 copies of each H2A, H2B, H3 and H4 and binds ~147 base pairs, which corresponds to ~1.65 superhelical turns (Luger et al., 1997). Additionally, the linker histone H1 can bind nucleosomes, which causes the nucleosome to bind a longer stretch of DNA, and stabilizes the overall DNA-nucleosome interaction (reviewed in Woodcock et al., 2006). All histones are highly basic proteins, and the histone octamer binds DNA essentially independent of sequence through interactions with the acidic phosphate groups. Histone proteins in chromatin are heavily post-translationally modified (reviewed in e.g. Andrews et al., 2016). The most studied modifications are lysine acetylation or methylation marks. Some of the acetylation marks probably affect the chromatin structure directly through negating the positive charge of the amino acid and disrupting the interactions of

nucleosome with DNA (or proteins, including other histones). For example, acetylation of lysines in the nucleosome core, such as H3K64ac or H3K122ac, as opposed to the histone tails, directly decreases nucleosome stability promoting transcription *in vitro* and in cells (Di Cerbo et al., 2014; Pradeepa et al., 2016; Tropberger et al., 2013). Other marks, however, require special protein readers that exert their function when binding to the modified nucleosomes.

Except for histone modifications, another layer of regulation of chromatin structure and function are histone variants: different histone protein paralogues that can be incorporated into the nucleosome instead of canonical histones (reviewed in Henikoff and Smith, 2015). Histone variants have a variety of functions. For example, one of the most well studied variants is CENP-A (or cenH3), an H3 variant, which is a key determinant of the centromere in eukaryotes. H2A histone has a particularly high number of variants. H2A.X is a well-known protein found across the genome and involved in DNA damage signalling: its phosphorylation produces ɣH2A.X, this modification spreads around the damaged site and recruits repair factors. H2A.Z is another important highly diverged H2A variant: nucleosomes containing H2A.Z are mostly found next to the gene promoters. These nucleosomes are often more labile, in particular when H2A.Z is found together with H3.3, an H3 variant found at active chromatin. Interestingly, like Polycomb proteins (see 1.3) H2A.Z is generally found at CpG islands (Illingworth et al., 2012) (see 1.2.1), in particular at the +1 nucleosome after the transcription start sites (TSSs).

Chromatin structure above the 10 nm fibre remains difficult to study, and there is no consensus regarding presence of a specific level of organization beyond that of nucleosomes. A popular model is that the 10 nm fibre folds to form the 30 nm fibre, usually believed to be a two-start helix, for example described using cryo-EM (Song et al., 2014). However all reports of 30 nm fibre structures have been made *in vitro*, after releasing chromatin from nuclei or assembling chromosomes on an artificial template (Razin and Gavrilov, 2014). Moreover the exact configuration of nucleosomes in the 30 nm fibre is highly dependant on the conditions used during the experiments: ion concentrations, presence of linker histones and nucleosome repeat length all affect the reported structures. Studying chromatin structure *in situ* is challenging due to very high density of nucleosomes and the small size of structures of interest. However, several studies using variations of electron microscopy have been reported. Using electron spectroscopic imaging (ESI) combined with electron tomography Fussner and colleagues have reported exclusively 10 nm fibres to be present *in situ* in the chromocentres and surrounding chromatin of mouse embryonic fibroblasts, as well as a variety of mouse tissues. They did not observe any higher order structures, while they were able to detect 30 nm fibres in starfish sperm nuclei, confirming the method did not disrupt them (Fussner et al., 2012). More recently, Ou and colleagues developed chromEMT, a method to selectively stain DNA for electron microscopy, and investigated chromatin structure in human small-airway epithelial cells (Ou et al., 2017). They observed a variety of structures with varying diameters mostly between 5 and 24 nm, however no regular structure was observed. Therefore, according to

electron microscopy, the 30 nm fibre probably can only be found in very special environments (e. g. sperm chromatin), or is a sample preparation artefact caused by chromatin dilution or removal of some endogenous components.

Lastly, two reports of high resolution micro-C (Hi-C using micrococcal nuclease for chromatin fragmentation, see 1.1.2.2 for details), shared a few months ago, also addressed this question genome-wide, using human HFF-1 and mouse ES cells, respectively (Hsieh et al., 2019; Krietenstein et al., 2019). Since micro-C analyses interactions between nucleosomes, this method is well suited for assessing the structure of the chromatin fibre. Interestingly, in both studies authors found a signature of a 2-start helical 30 nm fibre by quantifying interactions with increasingly linearly distal nucleosomes, and in particular by demonstrating similar interaction frequencies between N/N+2i and N/N+2i+1 nucleosomes (Figure 1.1).

While presented in different ways, results of the two papers look very similar and suggest presence of 30 nm fibre-like structures in mammalian genomes containing “small (~3-10) “clutches” of nucleosomes” (Krietenstein et al., 2019) or “2-3 tetra-nucleosome stacks” (Hsieh et al., 2019), however, their abundance and genomic distribution were not addressed, and it’s precise structure is difficult to estimate from these data.

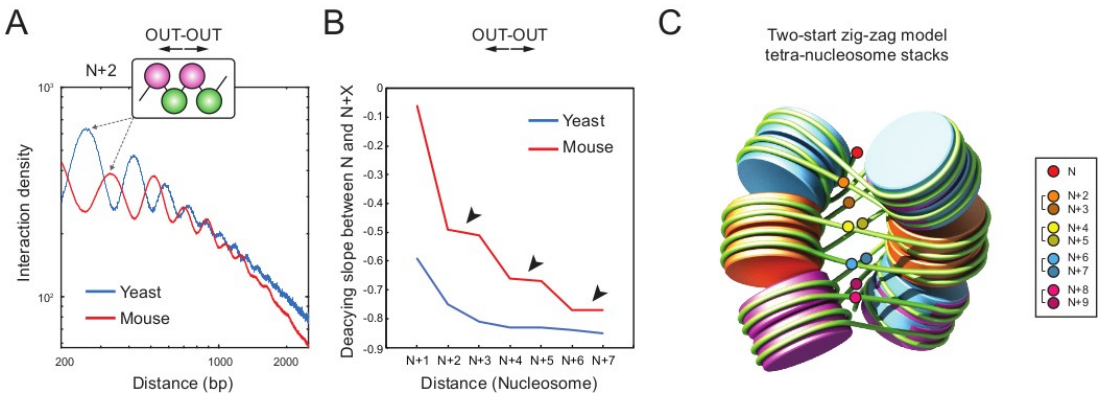

Figure 1.1. 2-start helical 30 nm fibre signal in micro-C data. Figure from (Hsieh et al., 2019). **(A)** Decay of interactions with distance in micro-C data from mouse and yeast cells for out-facing read pairs. The peaks of interaction frequency correspond to consecutive nucleosomes. **(B)** Quantification of decay of contact frequency between N and N+X nucleocomes. The curves show the slope of lines drawn between the peaks in (A). **(C)** Model of a two-start helical 30 nm fibre with colour-coded circles denoting out-facing interaction partners for each nucleosome.

1.1.2 Large scale chromatin organization, and methods to study it

The level of sub-nucleosomal organization and of the 10 nm fibre, i.e. core structure of the nucleosome, histone variants and modifications, is well understood; how nucleosomes assemble on the next level, such as a 30 nm fibre, and whether such a level even exists, is a question of debates in the field. However, the research into even higher order folding of the chromatin (at multi-kbp to Mbp scale) benefits from light microscopy methods such as DNA Fluorescence *in situ* Hybridization (DNA FISH), and also received a huge boost recently upon the advent of high throughput sequencing and chromosome conformation capture-methods, such as Hi-C used in this study.

1.1.2.1 FISH

FISH employs fluorescent probes complementary to specific sequences in the genome to visualize their location in the cell, either in an interphase nucleus, or in metaphase chromosomes. It requires fixation of the sample (usually, using paraformaldehyde in the case of interphase nuclei), and denaturation of the genomic DNA to allow its hybridisation to the probes. Probes can either be indirectly labelled with a moiety that can be easily detected (such as biotin), or directly labelled with fluorescent nucleotides. The former approach generates brighter signal, but requires additional steps, and can also increase the non-specific fluorescence. Probes are commonly generated using nick translation or other techniques from BAC or fosmid DNA, taking advantage of big BAC/fosmid libraries available for model organisms (Kim et al., 1992; Shizuya et al., 1992). Alternatively, oligonucleotide probe pools can be designed *in silico* and then synthesized (Beliveau et al., 2012; Boyle et al., 2011; Yamada et al., 2011); this approach based on oligonucleotides with is more expensive, but allows to avoid repeat regions and generate probes for any regions of any size.

FISH has been widely used to study the 3D organization of the genome. Simplest applications of FISH require labelling a single region in the genome, which can then be analyzed relative to nuclear landmarks, for example, to measure their radial position: in these early FISH experiments it was shown that proximity to the periphery (later termed Lamina Associated Domains - LADs) is one of the hallmarks of repressed chromatin, consistently with presence of heterochromatin at the edge of the nucleus close to the nuclear

lamina (Fawcett, 1966) even at whole-chromosome level (Boyle et al., 2001; Cremer et al., 2001, 2003; Croft et al., 1999). If one labels 2 regions in the genome using different fluorophores, it's possible to measure the distance between them. This distance can be a proxy for compaction of chromatin between these probes, if they are located close together (for example, Eskeland et al., 2010; Therizols et al., 2014). Alternatively, this distance can indicate looping between distal loci, often correlated with enhancer-promoter communication (for example, Williamson et al., 2016). Recent advances in FISH methods allow imaging dozens and even hundreds of sequences in the genome, which makes it possible to trace the path of a chromatin fibre in space for extended regions, and even quantify RNA molecules in the same cells to correlate the conformation of DNA and transcriptional status, however they require a complicated microscopic set-up and remain inaccessible to most researchers (Bintu et al., 2018; Cardozo Gizzi et al., 2019; Mateo et al., 2019; Nir et al., 2018; Wang et al., 2016). At the same time, advances in super-resolution imaging techniques, such as STORM and SIM, have allowed interrogation of more fine structures and probes in closer proximity (Boettiger et al., 2016; Fabre et al., 2015; Kundu et al., 2017; Smeets et al., 2014).

1.1.2.2 Hi-C and other C-methods

C-methods are a family of molecular approaches to map chromatin contacts. Starting from the simplest Chromosome Conformation Capture (3C) (Dekker et al., 2002), they all rely on the same principles (Figure 1.2). Usually, the cells are crosslinked using formaldehyde, then permeabilised, and the cytoplasm is removed. Then the chromatin is fragmented *in situ* (Gavrilov et al., 2013) using

a restriction enzyme (with alternative approaches using DNAse I (Ma et al., 2015) or micrococcal nuclease (Hsieh et al., 2015, 2016). Then the DNA ends are ligated using a T4 DNA ligase. Often the ends would not ligate back together with their original neighbouring restriction fragment, but with another one that is in close proximity in 3D space. These distal contacts are then quantified using different approaches (qPCR to quantify interactions between specific fragments in 3C, secondary digestion and circularization followed by inverse PCR and sequencing to quantify all interactions of a single fragment in 4C (Simonis et al., 2006; Zhao et al., 2006), oligo annealing and Ligation-Mediated Amplification (LMA) followed by sequencing to quantify all pairwise interactions between a set of fragments in 5C (Dostie et al., 2006), and paired end deep sequencing to measure the contacts genome-wide in Hi-C (Lieberman-Aiden et al., 2009). Of the commonly used C-methods, only Hi-C involves additional steps in the core 3C protocol: the ends of restriction fragments are filled in to incorporate a biotin moiety before they are ligated, which is then used for pull-down on streptavidin beads before sequencing to enrich for ligation products (Figure 1.2) (Lieberman-Aiden et al., 2009).

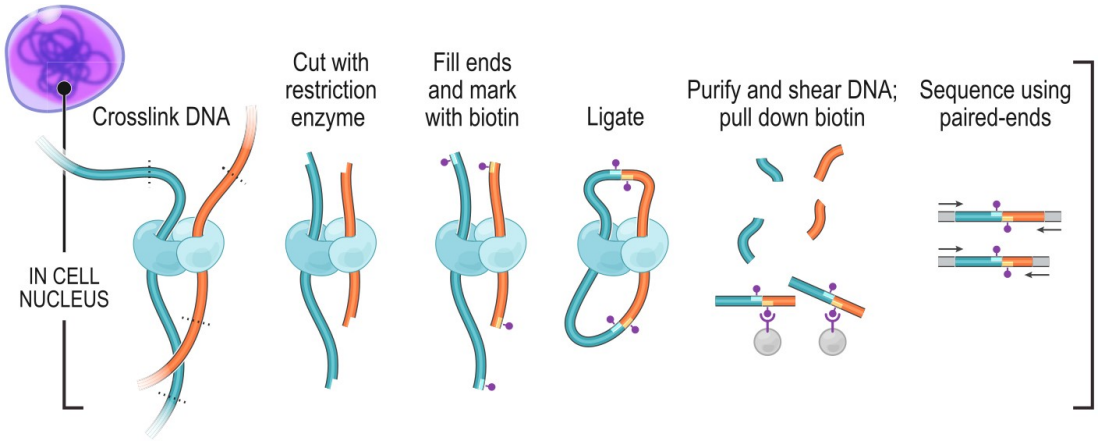

Figure 1.2. Schematic of Hi-C. Cells are first crosslinked using formaldehyde, then the DNA in the nuclei is fragmented using a restriction enzyme, and ends are filled and marked with biotin. Action of DNA ligase then stitches nearby ends together. DNA is isolated, fragmented, and sequenced after biotin pull-down. Schematic from (Rao et al., 2014).

Hi-C is the most powerful of C-methods, because it probes the chromatin organization genome-wide ("all-vs-all"), and the resolution is limited only by the sequencing depth (and restriction fragment size, but using frequent cutter enzymes, such as DpnII/MboI with average fragment size of ~450 bp, ensures that only extremely deep sequencing can cause cutting frequency to be limiting). However, due to the large number of potential interactions between all genomic regions, it requires very deep sequencing relative to most genomic methods to investigate 3D chromatin structure at high resolution (Lajoie et al., 2015). While C-methods, and Hi-C in particular, are very powerful due to their high throughput and high resolution, there are some important limitations that need to be mentioned. First of all, it is unclear what exactly a "contact" in Hi-C means: how close to each other in 3D do the fragments need to be to ligate in a Hi-C experiment? It probably depends on the protocol, for example, on the

choice of the restriction enzyme: longer fragments can reach further and can, presumably, ligate with more distal regions. Similarly, the cross-linking protocol (concentration of formaldehyde, cross-linking time) can affect the data properties (Flyamer et al., 2017). The next issue, common to all bulk methods, is that they reflect a population average of a large number of cells. While there are approaches to perform Hi-C in single cells (Flyamer et al., 2017; Nagano et al., 2013, 2017; Ramani et al., 2017; Stevens et al., 2017; Tan et al., 2018), they are complicated and don't provide enough data for robust high-resolution analysis even in combined data from multiple cells. Interestingly, DNA FISH is complementary to C-methods: their throughput and resolution are limited, but the measurements are simple physical distances, and they come from individual cells. Both of these groups of methods, however, can only be applied to fixed cells, and lack any information about chromatin dynamics in time.

1.1.2.3 Hi-C data analysis

Like many high throughput sequencing methods, Hi-C requires complicated and multistage computational analysis to make sense of the data. First, the reads need to be mapped to the reference genome. Since the distance between paired reads in a Hi-C library can be very long, usually the two ends have to be mapped separately. The ligation junction is often sequenced through from one or both of the sides, which would cause many mapping approaches to fail on these "chimeric" reads, because the read consists of two potentially very distal genomic fragments (Imakaev et al., 2012). Before development of fast and accurate local alignment methods (Li, 2013), an iterative mapping approach could be used to mitigate this complication: by first

mapping short sequences from the ends of reads and iteratively extending them for unmapped reads allowed to use chimeric reads (Imakaev et al., 2012). However, using local mapping and thus rescuing chimeric reads allows much faster processing without any loss of accuracy (Rao et al., 2014; Servant et al., 2015).

Then the mapped reads need to be filtered. Some filters are essential (removing single-side mapped reads, or PCR/optical duplicates), while others are often implemented, but not really required (removing reads from the same restrictions fragment, or other invalid or not useful ligation products). PCR/optical duplicates can not be removed downstream, however all other artefacts can be very simply filtered out by discounting contacts within a certain distance (Ma et al., 2015), while they might be incompletely removed by more complicated filters, and ignoring very short-range contacts is standard practice anyway. Another filtering step that has to be done at this stage is removing multi-mapping or poorly mapped reads.

For most analyses, filtered reads (which are also called “contacts”) are then binned into a matrix of specified resolution (also known as a Hi-C map). For example, a 5 kbp resolution matrix means the size of each genomic bin is 5,000 bp, and each value in the matrix contains the number of contacts that connect the two sides of this pixel. There is no rule for choosing the right resolution for a particular dataset; moreover, different analyses are often performed at different resolutions. Generally, using smaller bin size will show more fine details in the Hi-C maps, but noise level will also be higher (Lajoie et al., 2015). It is often

beneficial to generate Hi-C maps at different resolutions, and then visually explore them to choose the most appropriate bin size (see Figure 1.3).

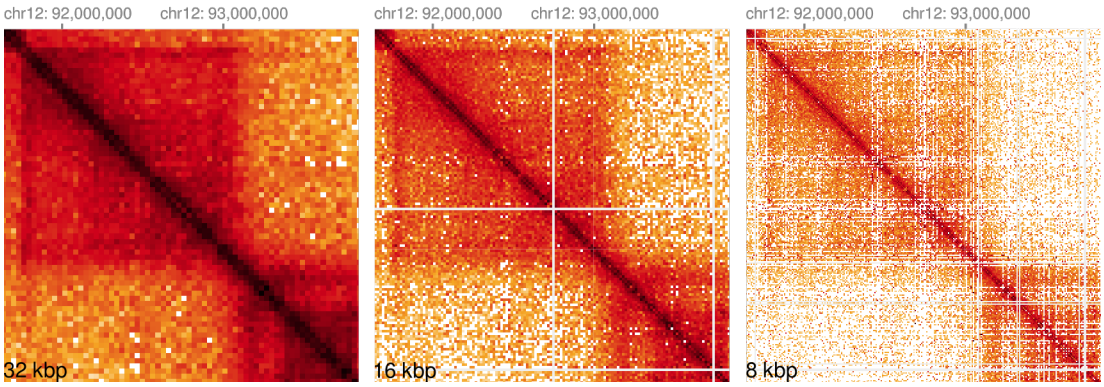

Figure 1.3. Hi-C contact maps at different resolutions. The same region from chromosome 12 in Hi-C data from human ES cells (Dixon et al., 2012) at three different resolutions. While 32 kbp and 16 kbp are acceptable to use, 8 kbp bins are too small: data is very noisy and a lot of bins are blank. Generated using HiGlass.

Raw Hi-C data is biased: certain regions have artificially high or low level of interactions. This can mostly be attributed to restriction site frequency and fragment length, GC content and mappability, and one of the approaches to mitigate the biases is by directly modelling their effects on Hi-C data, and removing them (Yaffe and Tanay, 2011). An alternative approach, agnostic to the source of biases, is iterative correction, or matrix balancing (Imakaev et al., 2012). This method assumes that all genomic regions should have an equal "visibility", or total number of contacts with other regions, and that all biases are factorizable, i.e. that biases of two distal regions are multiplied for the contact frequency observed between them. For each genomic bin, it finds the weight scores that, when used for multiplication of contact frequencies originating from them, produce a uniform coverage across all bins. It has been shown that, at least at low resolution, both approaches produce near identical results,

however the second method does not make assumptions about the nature of biases present in the data, and is much more computationally efficient (Imakaev et al., 2012). All further analyses of Hi-C data usually use balanced (or otherwise corrected) data.

1.1.2.4 Finding patterns in Hi-C data

When looking at the whole-genome, the most striking features are the dark squares, with high contact frequency inside them, which correspond to interactions within the chromosomes (Figure 1.4). This indicates the existence of chromosome territories: during interphase, each chromosome occupies a defined volume in the nucleus, and different chromosomes don't significantly intermingle. They have been studied extensively by FISH using "chromosome paints" – pools of FISH probes covering whole chromosomes (reviewed in Cremer and Cremer, 2010). The strength of chromosome territory separation can be quantified using *cis*/*trans* contact fraction: the more *cis*, and the fewer *trans* contacts are found in Hi-C data, the more chromosomes are separated.

Another major features found in Hi-C data is the apparent plaid pattern observed at relatively low resolution (hundreds of kbp – low Mbp), which corresponds to nuclear compartments: A (active) and B (inactive), and regions within each of the compartments preferentially interact with other regions within the same compartment (Lieberman-Aiden et al., 2009) (Figure 1.5A). This observation is consistent with earlier FISH analysis showing spatial segregation of gene-rich and gene-poor regions in the mouse nucleus (Shopland et al., 2006). Compartmentalization in Hi-C data can be captured by the first eigenvector (same as the first principal component) of the balanced

data, and this pattern correlates with chromatin activity, such as histone marks, gene expression, replication timing and Dnase I accessibility, and GC content which is indicative of gene density (Imakaev et al., 2012; Lieberman-Aiden et al., 2009).

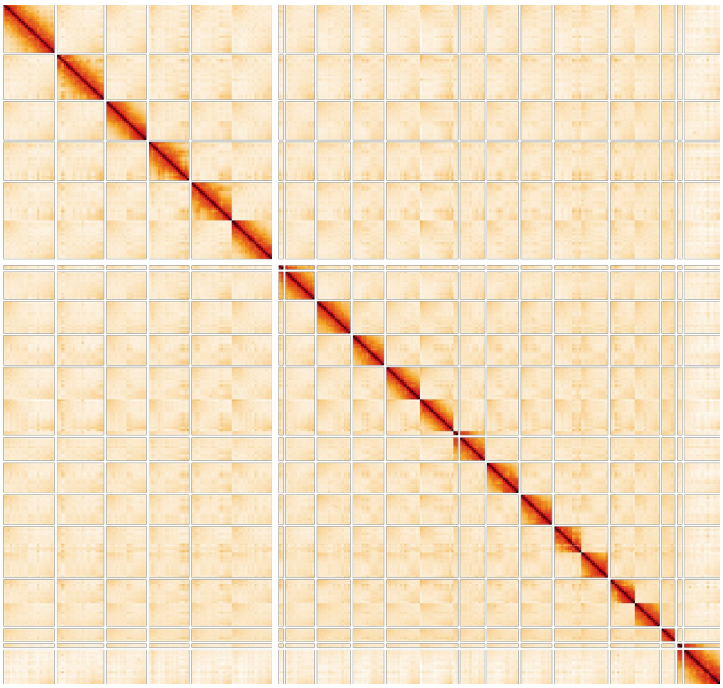

Figure 1.4. Whole-genome Hi-C map of mouse ES cells at 8.192 Mbp resolution. Data from (Nora et al., 2017). Darker colours indicate higher contact frequency. All chromosomes are shown, from chr1 (top left) to chrX (bottom right). Signal from chromosome territories is visible as dark squares along the main diagonal.

A/B-compartmentalization corresponds to the radial organization of the nucleus, and separation of heterochromatin from euchromatin: B compartment mostly consists of Lamina-Associated Domains (LADs) and Nucleolus-Associated Domains (NADs), while the A-compartment consists of active

chromatin found away from nuclear periphery and nucleoli (Chen et al., 2018; Stevens et al., 2017). This has been confirmed genome-wide using Tyramide signal amplification followed by sequencing (TSA-seq): an approach to measure distance of genomic regions from nuclear cytological compartments using antibody-HRP (horse radish preoxidase) fusion, and addition of tyramide-biotin to fixed cells (Chen et al., 2018). Diffusing tyramide-biotin radicals then function as a molecular ruler to estimate average distance of genomic loci from the protein of interest. Application of TSA-seq to nuclear speckles and lamina revealed a pattern of radial organization very similar to that uncovered by Hi-C.

More fine Hi-C compartmentalization analysis using clustering reveals that both A and B compartments can be further subdivided into sub-compartments (Rao et al., 2014). Of the two A-subcompartments, A1 is associated with nuclear speckles (nuclear bodies involved in splicing, RNA metabolism and transcriptional regulation (reviewed in Galganski et al., 2017)) and is more active than A2 (Chen et al., 2018; Rao et al., 2014). Similarly, B-subcompartments are associated with different histone marks and flavours of heterochromatin: B1 is enriched in H3K27me3, B2 – in pericentromeric heterochromatin and NADs, and B3 – in LADs, while B4 is very small and includes the clusters of KRAB-ZNF genes (Rao et al., 2014). Interestingly, according to previous reports (Vieux-Rochas et al., 2015) and my observations (data not shown), unlike the above analysis performed in GM12878 cells, in mouse ES cells H3K27me3 is associated with A compartment. The formation of such compartments in general is probably associated with (micro)phase

separation and like-like association of regions of the chromatin fibre (Falk et al., 2019; Mirny et al., 2019).

Microphase separation of chromatin and liquid-liquid phase separation of nucleoplasm components have been proposed as a key mechanism driving segregation of different nuclear compartments (reviewed in Sawyer et al., 2019). These are physical phenomena of “oil in water”-like behaviour: either blocks of a polymer, or different components of a solution, spatially segregate according to their mutual attraction. The most well known example of this behaviour is the nucleolus: a separate nuclear compartment containing much lower concentration of DNA than surrounding chromatin, but with high levels of RNA and associated proteins. Similarly, heterochromatin has been shown to form a separate phase in the nucleus due to HP1, one of the main heterochromatin components (Larson et al., 2017; Strom et al., 2017). Together with anchoring of heterochromatin at the nuclear periphery, this leads to a radial organization of the nucleus with heterochromatin primarily found on the edge, and more active compartment inside. Abolishing lamin-based peripheral anchoring of heterochromatin does not lead to intermixing of active and inactive regions due to self-attraction of heterochromatin, and these nuclei become “inverted”: heterochromatin forms a compartment in the centre of the nucleus instead of its periphery (Falk et al., 2019).

When investigating Hi-C maps at higher resolution, regions of enriched interaction frequencies on the scale of 100 kbp – 1 Mbp close to the main diagonal are observed (Figure 1.5B) (Dixon et al., 2012; Nora et al., 2012; Rao et al., 2014). They were termed Topologically Associating Domains (TADs) and

have been a focus of a lot of research in recent years. There is a multitude of methods used for their annotation, but no one algorithm is considered a gold standard (Forcato et al., 2017; Zufferey et al., 2018). The original approach relied on Directionality Index (DI): a ratio of short-range interactions upstream and downstream for each genomic bin (Dixon et al., 2012). Much more complicated methods have been developed recently, including approaches that detect hierarchical TAD structures (Wang et al., 2017b; Weinreb and Raphael, 2016). A very simple a robust way of finding TAD boundaries, however, is simple Insulation Score (IS), defined as the number of contacts in a diamond-shaped window above the genomic bin of interest: valleys of this measurement correspond to high insulation, and detecting them allows finding TAD boundaries easily (Crane et al., 2015).

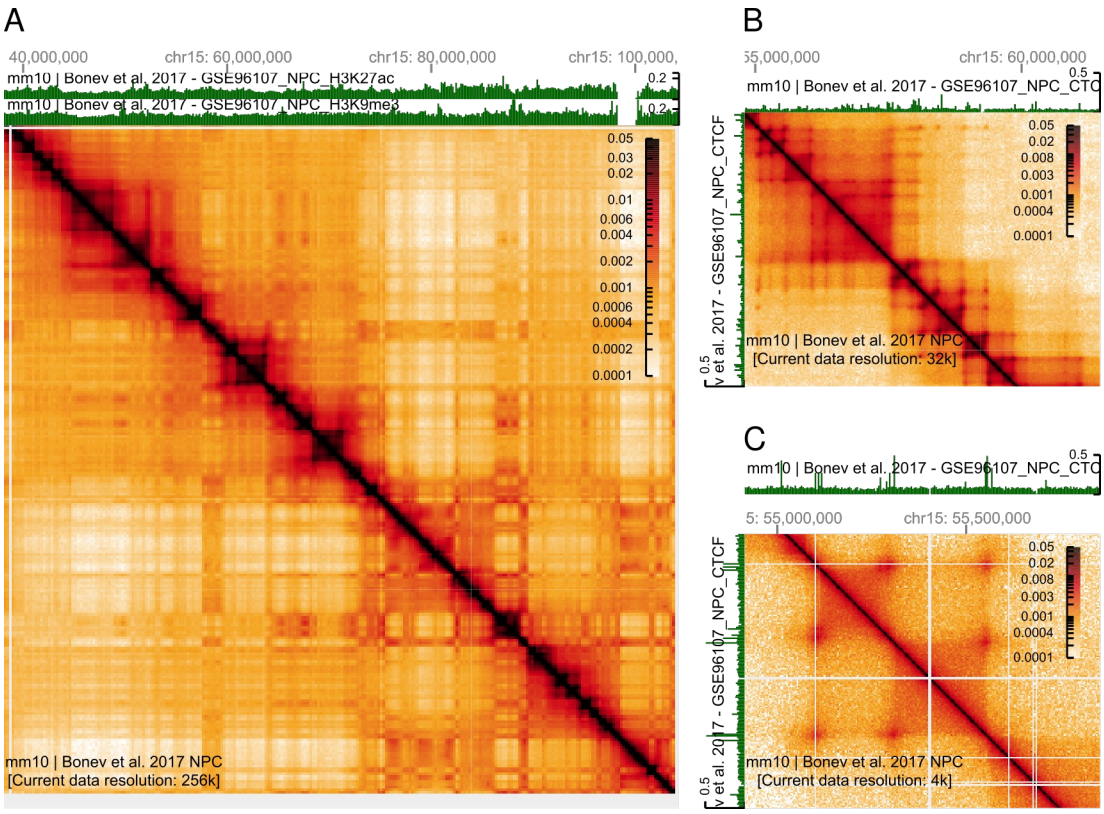

Figure 1.5. Features observed in Hi-C maps. (A) Compartments. A part of chromosome 15 from Neural Progenitor Cells (NPC) is shown: Hi-C map at 256 kbp resolution with H3K27ac and H3K9me3 ChIP-seq track above (the data here and in other panels is from Bonev et al., 2017)(data here and in other panels from Bonev et al., 2017). All panels created using HiGlass. (B) TADs. A small part of the Hi-C map from (A) is shown, together with a CTCF ChIP-seq track. (C) Loops. A part of the large TAD from (B) is shown, together with a CTCF ChIP-seq track.

After the discovery of chromatin loops in high-resolution mammalian Hi-C data (Figure 1.5C) the mechanism of TAD formation has been elucidated (Rao et al., 2014). Loops were found using an algorithm called HiCCUPS that ensured enrichment of interactions relative to local background and used a special modification of multiple testing correction based on “λ-chunking” - a procedure that takes the expected values for each loop into account. Loops frequently connected two ends of the TAD (or distal ends of neighbouring TADs), and the

vast majority of these loops had convergent CTCF binding sites, together with cohesin ChIP-seq peaks. CTCF is a protein that binds DNA sequence-specifically using its zinc-finger domains. It has long been implicated in structural organization of chromatin (reviewed in Ghirlando and Felsenfeld, 2016), and in particular it has been shown that CTCF binding sites have an insulator (enhancer blocking) activity (Bell et al., 1999). These properties of chromatin loops lead to the model of loop extrusion, whereby cohesin molecules create loops on chromatin fibers and can processively extrude them, until encountering a CTCF site in the inwards orientation, where it stops (Figure 1.6) (Fudenberg et al., 2016; Sanborn et al., 2015). In population-average Hi-C data this process leads to observed enrichment of interactions within TADs, and a prominent peak of interaction frequency between two CTCF sites. Direct *in vitro* evidence for the loop extrusion process by human cohesin was recently reported, and shown to require NIPBL/MAU2 cofactors (Davidson et al., 2019). Similarly, the molecular mechanism of cohesin-CTCF interaction has recently been elucidated: N-terminal domain of CTCF binds cohesin through SCC1/SA2 subunits; this interaction also stabilizes cohesin on chromatin by competeing with the cohesin removal factor WAPL, that acts through the same protein surface (Li et al., 2020; Nora et al., 2019; Pugacheva et al., 2020).

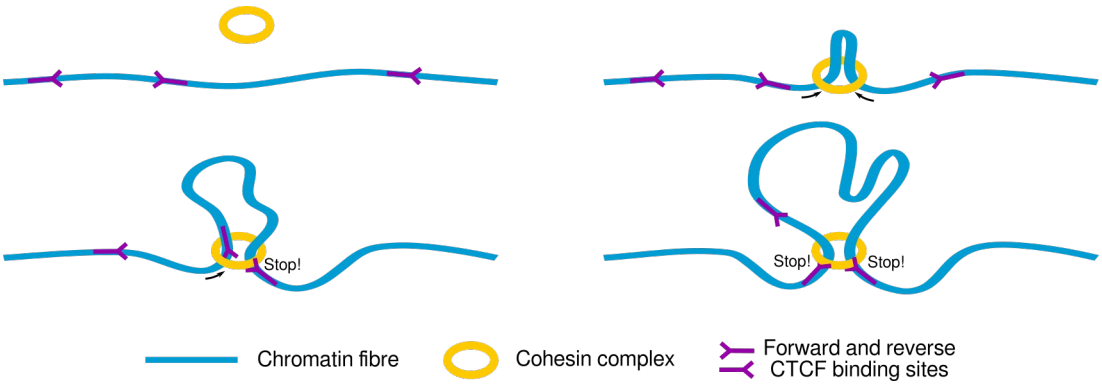

Figure 1.6. The loop extrusion model. The extrusion complex consists of one or more cohesin molecules. CTCF motifs in the inwards orientation, when bound by CTCF, stop the extruding complex, while sites in the opposite orientation allow cohesin to extrude through them.

While loop extrusion can have far-reaching consequences for genome and chromatin biology, demonstrating its biological functions has been challenging so far. One of the favourite candidates has been mediation of enhancer-promoter interactions, and regulation of gene expression. However upon efficient depletion of CTCF or cohesin and following almost complete loss of TADs and loops, gene expression is only very mildly perturbed (Nora et al., 2017; Rao et al., 2017b). In contrast, local perturbations of TADs has been reported to affect gene regulation by enhancers (Despang et al., 2019; Franke et al., 2016; Lupiáñez et al., 2015), although not in every studied case (Paliou et al., 2019; Williamson et al., 2019). Another function for this process could be in un-knotting the genome. It has been reported that etoposide-induced topoisomerase II mediated breaks are strongly associated with loop anchors, and their positions match closely the CTCF and cohesin binding sites (Canela et al., 2017, 2019; Gothe et al., 2019). This leads to the model whereby cohesin

untangles the genome during loop extrusion, and the accumulated topological problems are resolved by topoisomerase II at loop boundaries.

Interestingly, loop extrusion and chromatin compartmentalization (including polycomb-mediated distal interactions and superenhancer clustering) coexist, but the former interferes with the latter: removal of cohesin accentuates compartments and other like-like distal interactions (Mirny et al., 2019; Rao et al., 2017a; Rhodes et al., 2020; Schwarzer et al., 2017). Possibly, loop extrusion acts as a dynamic “lubricant” that prevents phase separation progressing too far and keeps the chromatin dynamic.

1.2 DNA methylation

The most widespread chemical modification of DNA bases found in mammalian genomes is methylation of cytosine at the 5th position of the ring in a CpG context (and modifications derived from it during demethylation). The majority of CG dinucleotides are methylated in the genome (estimated at ~80% (Bird, 2011)), although this highly depends on the cell type. This is a repressive modification that is deposited by DNA methyltransferases (DNMTs; see 1.2.2, and (reviewed in Lyko, 2018)). DNA methylation is a mark found throughout the genome (Figure 1.7). It was proposed to have epigenetic roles a long time ago (Holliday and Pugh, 1975; Riggs, 1975), however uncovering its direct role in gene regulation has been challenging, despite early work showing transcriptional inhibition (Vardimon et al., 1982). It appears, DNA methylation is primarily used to silence transposable elements, as well as imprinting, X chromosome inactivation, and regulation of some enhancers and a small subset of gene promoters (Edwards et al., 2017; Hackett et al., 2012; Jones, 2012).

Figure 1.7. DNA methylation landscape of mammalian genomes. Schematic depiction of patterns of CpG methylation in some genomic contexts.

1.2.1 CpG islands

5mC has mutagenic potential due to spontaneous deamination to thymidine (Shen et al., 1994) which is often erroneously repaired to T-A instead of C-G (Hendrich et al., 1999; Waters and Swann, 2000), and therefore overall the genome is depleted of CG dinucleotides. However some regions termed CpG islands (CGIs) contain unusually long stretches of CpGs with low level of their CpG methylation (Bird et al., 1985; Cooper et al., 1983; Gardiner-Garden and Frommer, 1987).

CGIs are often associated with gene promoters, in particular the majority of housekeeping genes use CGI promoters (Lander et al., 2001; Larsen et al., 1992; Zhu et al., 2008). While some CGIs do not localize in the proximity of an annotated transcription start site (TSS), they might act as highly tissue specific promoters that escaped annotation (Illingworth and Bird, 2009). The majority of CGIs are never methylated in healthy conditions, however a small subset of them acquires DNA methylation during normal development. The most important examples of these are some genes involved in germ cell development, X chromosome inactivation and genomic imprinting (Edwards and Ferguson-Smith, 2007; Reik, 2007). Interestingly, methylation of some CGIs together with other perturbation of the methylome frequently occurs in various cancers (reviewed in Klutstein et al., 2016). It is however unclear how CpG islands normally avoid acquiring DNA methylation. Different possibilities have been discussed, but the likely candidates involve DNMT activity inhibition of some sort at CGIs: either steric hindrance from binding of transcription factors and transcription machinery during establishment of DNA methylation

patterns during development, or refractory properties of CGI chromatin to action of DNMTs, such as histone modifications (Illingworth and Bird, 2009). It is clear, however, that the hypomethylated property of CGIs is sequence-dependent, and the mechanisms are conserved across human, mouse and zebrafish (Long et al., 2016).

1.2.2 DNA methylation enzymes

The 5mC mark is deposited by DNA methyltransferases (DNMTs), a family of enzymes containing a conserved catalytic domain that includes DNMT1, DNMT2, DNMT3A, DNMT3B, DNMT3C and DNMT3L in the mouse genome (reviewed in Lyko, 2018). DNMT2 turned out to be an RNA methyltransferase, unlike its relatives, and DNMT3L is a catalytically inactive protein, while the other members of the family can deposit DNA methylation. DNMT3C is a recently discovered enzyme specific to the male germline used to silence repetitive elements, and will not be discussed here (Barau et al., 2016).

1.2.2.1 De novo DNMTs

DNMT3A and DNMT3B are often together called the *de novo* DNMTs, in contrast with the maintenance DNMT1, since they can deposit DNA methylation using an unmethylated DNA template (Okano et al., 1999). These DNMTs establish methylation patterns during development, which can then be perpetuated by the maintenance DNMT1. These enzymes are highly expressed in early development, but are downregulated in differentiated cells (Okano et al., 1999). DNMT3L contains a truncated catalytic domain and is therefore enzymatically inactive, however it functions as a key cofactor for DNMT3A/B to

stimulate their activity (Chédin et al., 2002; Chen et al., 2005; Gowher et al., 2005; Hata et al., 2002; Suetake et al., 2004).

1.2.2.2 Maintenance DNMT – DNMT1

DNA methylation patterns, once established, can be propagated throughout cell divisions. DNMT1 is the enzyme that specifically binds hemimethylated CpGs, and catalyzes methylation of the unmethylated cytosin (Bestor and Ingram, 1983; Gruenbaum et al., 1982). Additionally, DNMT1 interacts with PCNA (Proliferating Cells Nuclear Antigen), a molecular clamp that follows the replication fork (Chuang et al., 1997; Iida et al., 2002). This interaction ensures that DNMT1 screens the newly synthesized DNA strand for hemimethylated CpG to faithfully propagate DNA methylation patterns. The key cofactor of DNMT1 is UHRF1, which can recruit DNMT1 to hemimethylated DNA, and also can target it to H3K9me3-containing chromatin (Bostick et al., 2007; Rothbart et al., 2012).

While this model of separation of maintenance and *de novo* functions of DNMTs is attractive in its simplicity, it is not entirely accurate (Jones and Liang, 2009). For example, ES cells lacking DNMT3A and DNMT3B, but with intact DNMT1, partially lose DNA methylation in certain regions, suggesting DNMT1 alone is not able to fully maintain 5mC levels (Chen et al., 2003; Liang et al., 2002; Okano et al., 1999). Moreover, DNMT1 can perform *de novo* DNA methylation after stimulation by presence of methyl groups on the DNA (Fatemi et al., 2002). Therefore these enzymes can cooperate with each other and probably at least partially perform both functions.

1.2.2.3 DNA demethylation

Loss of DNA methylation can occur in multiple ways. First, simple passive dilution of the methyl mark with cell division and low DNMT activity would efficiently deplete this mark. Alternatively, active DNA demethylation can occur via activity of TET (ten-eleven translocase) enzymes that remove the methyl mark by oxidation to 5-hydroxymethylcytosine (5hmC), 5-formylcytosine (5fC), and 5-carboxylcytosine (5caC), which can then be removed by base excision repair (reviewed in Ross and Bogdanovic, 2019). Both of these mechanisms have been shown to work in both primordial germ cells, and early mouse embryo (Guo et al., 2014; Kagiwada et al., 2013; Kawasaki et al., 2014; Ohno et al., 2013; Santos et al., 2013).

1.2.3 Mechanisms of action of DNA methylation

While associated with repression of transcription, it's been challenging to establish whether DNA methylation is actually causative in silencing (Schübeler, 2015). For example, engineered methylation of promoters causes repression of gene transcription (Korthauer and Irizarry, 2018). Similarly, DNA methylation affects at least some regions in a genome-wide episomal enhancer activity assay, although not always negatively (Lea et al., 2018). It is widely believed that methylation of CGIs causes silencing of the associated genes (Bird, 2002), but *in vivo* examples of CGIs that get methylated as part of gene regulation mechanisms are scarce. Examples include some germ line specific genes and other tissue specific genes, developmental genes, and the mostly well studied imprinted genes and the inactive X chromosome, but never more than a few percent of all genes in a given cell type have a methylated CGI

promoter (Illingworth and Bird, 2009). Interestingly, this differential methylation can affect only one of the alternative promoters, and this way can cause promoter switching (Rauch et al., 2009). As mentioned above, artificial deposition of 5mC mark at gene promoters causes gene repression (Korthauer and Irizarry, 2018). Therefore DNA methylation is able to silence genes, but whether, and how often, methylation actually pre-dates and causes silencing *in vivo*, in contrast to merely following on to another silencing event to "lock it down", is less clear. It is more clear, however, that DNA methylation is used to silence transposable elements (TEs) in a wide variety of organisms, and it has been proposed as its primary function (reviewed in Deniz et al., 2019).

How does methylation silence transcription? There are multiple ways for an epigenetic mark to exert its function. First, it can directly affect physical properties of underlying DNA or chromatin, and increased stiffness of hypermethylated DNA has been reported (Cassina et al., 2016; Onoshima et al., 2017). This can directly affect DNA-protein interactions. Similarly, the presence of the methyl mark itself within the recognized motif can influence binding of many transcription factors, both positively and negatively (Yin et al., 2017; Zuo et al., 2017).

Alternatively, 5mCG dinucleotides can be recognized by special reader proteins. Methyl-CpG binding domain (MBD) proteins, such as MeCP2 (methyl CpG binding protein 2), have a strong preference for methylated CpGs (Meehan et al., 1989). These proteins can, in turn, recruit specific silencing activities, such as histone methyltransferases and deacetylases (reviewed in Du et al., 2015). MBD proteins can have specific roles in different processes or

cell types, for example MeCP2 is particularly important in neural cells. Rett syndrome is, usually, caused by mutations in its MBD or TRD (transcriptional repression domain**)** domains.

Conversely, some proteins could specifically recognize non-methylated CG dinucleotides. These are the CXXC proteins that are capable of specific binding to unmethylated DNA, and therefore primarily recognize CGIs in the vertebrate genomes. For example, Cfp1 (CxxC finger protein 1) targets H3K4me3 mark to CGIs (Thomson et al., 2010), while KDM2B (lysine (K)-specific demethylase 2B) is a component of Polycomb repressive complex 1 and is one of its key recognizing activities to bind CGIs (Blackledge et al., 2014).

1.2.4 Dynamics of DNA methylation during development

While DNA methylation is universally high in adult somatic tissues with at least approximately 80% CpGs methylated (Ehrlich et al., 1982), its levels are much more variable during development and in the germ line (reviewed in Lee et al., 2014) (Figure 1.8). At the beginning of an organism's development, two gametes fuse to generate the one-cell embryo, or zygote. Sperm cells have very high levels of DNA methylation, while oocytes have a lower level. Interestingly, in the zygote two nuclei are separate throughout the cell cycle, but both genomes undergo demethylation. The loss of DNA methylation progresses throughout the first days of early mouse development, but starts rising again around E3.5 to reach high levels by the epiblast stage in E6.5. At this stage the primordial germ cells (PGCs) are specified, and they again undergo a wave of global demethylation, only to acquire it back in later stages of their development. In both waves of demethylation DNMT1 activity is impaired,

either by transcriptional downregulation of its cofactor UHRF1, or by excluding it from the nucleus (Kagiwada et al., 2013; Seisenberger et al., 2012), however the active DNA demethylation by TET enzymes also plays a role, in particular in the male pronucleus (Guo et al., 2014; Santos et al., 2013).

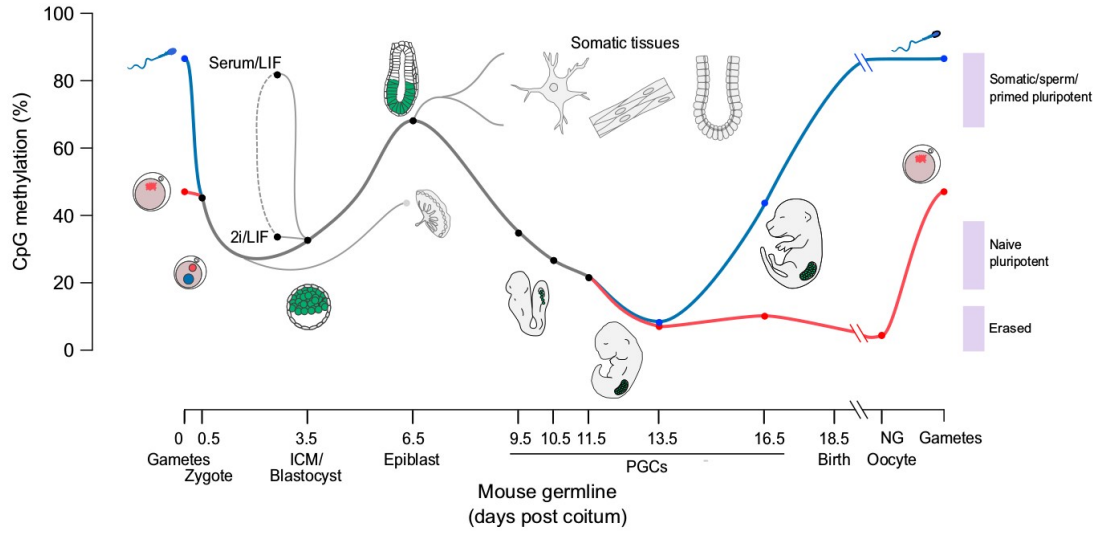

Figure 1.8. Global levels of DNA methylation across mouse development. DNA methylation level shown here at different stages of mouse development, in both somatic and germline cells. Also ES cells cultured in serum/LIF and 2i/LIF are shown (see below for details). From (Lee et al., 2014).

1.3 Polycomb Repressive Complexes

Polycomb Repressive Complexes (PRC) are a system of negative regulators of gene expression. They are crucial for correct development in a variety of multicellular organisms, since they silence developmental genes until their expression is required for a particular cell lineage. While discovered a long time ago in *Drosophila* (Lewis, 1947) as regulators of the homeobox gene expression (Lewis, 1978), this system has been receiving more and more attention over the years due to roles in diverse cellular functions on top of repression of developmental genes, from cell cycle control to cancer biology (Schuettengruber et al., 2017). There are two broad non-overlapping classes of PRCs, which often co-bind the genome and are both involved in gene repression: PRC1 and PRC2. All of their components are broadly called Polycomb-group (PcG) proteins.

1.3.1 PRC1

PRC1 complexes can modify chromatin using the core RING1A/B E3 Ubiquitin ligase subunit: they deposit the H2AK119ub mark at their target sites. Except for the RING subunit, the only other constitutively present binding partner in PRC1 complexes in one of six PCGF1-6 proteins, and which of the PCGF homologues takes part in forming the complex determines the PRC1 complex subtype (PRC1.1 to PRC1.6, according to the PCGF subunit) (Gao et al., 2012). Different PRC1 subtypes contain different subunits, however 1.2 and 1.4 have similar composition; same applies to 1.3 and 1.5 complexes (Figure 1.9). Additionally, PRC1 complexes can contain RYBP or YAF2, however the presence of these subunits is mutually exclusive with “canonical” components

of PRC1.2 and PRC1.4 due to both RYBP/YAF2 and CBX subunits binding RING1A/B through the same residues (Wang et al., 2010). Components of canonical PRC1 (cPRC1) are CBX, PHC and SCM proteins, homologues of components of PRC1 complexes originally described in *Drosophila* (Saurin et al., 2001; Shao et al., 1999) and observed in initial studies of mammalian PRC1 (Levine et al., 2002). Other subunits are termed non-canonical, and complexes containing them are non-canonical, or variant PRC1 – ncPRC1 or vPRC1. Interestingly, however, these non-canonical subunits co-purify together with PCGF2/4 (Gao et al., 2012), and therefore the PCGF subunit does not determine whether the complex is canonical or variant.

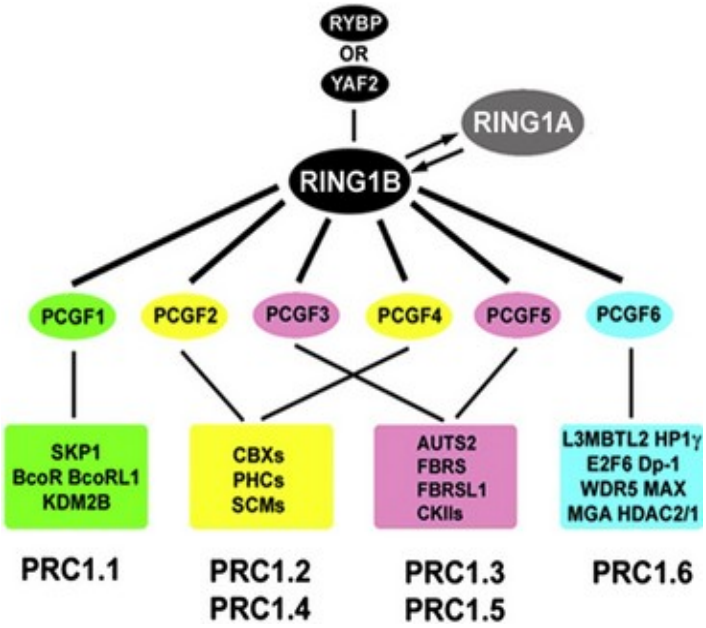

Figure 1.9. PRC1 subtypes. From (Gao et al., 2012).

1.3.1.1 Canonical PRC1

Canonical PRC1, or cPRC1, complexes contain CBX, PHC and SCM subunits.

CBX proteins are the homologues of *Drosophila* Polycomb (Pc) and contain chromodomains which bind methylated histones. Most of them bind

H3K27me3, the mark deposited by PRC2, however the level of specificity, in particular in distinguishing this mark from H3K9me3, varies between the CBX proteins (Bernstein et al., 2006; Kaustov et al., 2011). It has been proposed that this recognition of H3K27me3 is the main mechanism of PRC1 targeting to chromatin, following recruitment of PRC2 and deposition of this mark (Cao et al., 2002; Wang et al., 2004b). While this simple hierarchical model of PRC1 recruitment has recently been challenged (see 1.3.3), it probably functions to target specifically cPRC1 to regions of high H3K27me3 (Blackledge et al., 2014).

PHC subunits are homologues of *Drosophila* Polyhomeotic (Ph) and contain a Sterile Alpha Motif (SAM) domain. The SAM domain allows these proteins to oligomerize (reviewed in Kim and Bowie, 2003; Kim et al., 2002), which leads to their aggregation in the nucleus, and their participation in 3D organization of cPRC1 targets in the nucleus (Boettiger et al., 2016; Isono et al., 2013; Kundu et al., 2017; Wani et al., 2016).

Other sub-stoichometric components found in cPRC1 complexes are the SCM proteins, homologues of the Sex comb on midleg (Scm) protein in *Drosophila*. They are less studied, but they also contain SAM domains and bind to the PHC subunits in cPRC1 through them (Frey et al., 2016). It has been suggested that, in *Drosophila*, the Scm component can link PRC1 and PRC2 together, and coordinates their binding to chromatin (Kang et al., 2015).

1.3.1.2 Variant PRC1

PRC1 complexes that don't contain CBX, PHC and SCM subunits are called non-canonical, or variant. The only universal non-RING subunit of vPRC1 is the

YY1-binding protein (RYBP) (or its paralogue YY1-associated factor 2, YAF2) – worth noting, however, that despite the name, in mammalian cells YY1 does not bind PRC1, and doesn't co-bind the same sites on the genome (Gao et al., 2012; Mendenhall et al., 2010). RYBP has a non-specific DNA binding activity, but whether this has any function for PRC1 targeting has not been shown (Neira et al., 2009). In contrast, it has been shown to enhance the E3 Ubiquitin ligase activity of RING1B, suggesting vPRC1 to be the primary enzymatically active complexes – not cPRC1 (Rose et al., 2016; Tavares et al., 2012)

Some vPRC1 complexes contain other DNA binding proteins which could provide targeting specificity, and have been shown to direct vPRC1 to certain sites. The most studied example is KDM2B (lysine-specific demethylase 2B), which associates with PRC1.1. As an enzyme, it has histone demethylase activity specific to H3K36 and/or H3K4 (Frescas et al., 2007; Janzer et al., 2012; Tsukada et al., 2006). Additionally, using its CxxC domain, KDM2B binds CGIs (Koyama-Nasu et al., 2007). By artificial tethering to an ectopic genomic site via TetR-mediated recruitment , KDM2B has been shown to bring in the PRC1.1 complex which led to binding of PRC2 and both H2AK119ub and H3K27me3 modification of chromatin (Blackledge et al., 2014). This is consistent with the primary PRC1 targets in the genome being CGIs, however KDM2B binds all CGIs (He et al., 2013), but only a fraction of them is occupied by PRC1. It appears, that the explanation is simple: PRC complexes only bind CGIs in absence of transcription, and transcriptional inhibition is sufficient to induce PRC2 recruitment to nucleosome-free CGIs (Riising et al., 2014).

Other examples of DNA-binding proteins that take part in forming vPRC1 complexes (PRC1.6, in particular) are MYC-associated factor X (MAX) and its partner MAX gene-associated protein (MGA), which bind the MYC motif (Gao et al., 2012; Hurlin, 1999), and transcription factors E2F6 and DP1, which might target it to the E2F motif. However non-canonical PRC1 complexes are highly variable, and their detailed description is outside of the scope of this work.

1.3.2 PRC2

PRC2 is the other group of PRC complexes. It contains a core enzymatic subunit, EZH2 (Enhancer of zeste homologue 2, or a less efficient homologue EZH1 (Margueron et al., 2008)), which catalyzes addition of methyl groups to the H3K27 residue using its SET domain (Czermin et al., 2002; Müller et al., 2002). Other core subunits of PRC2 are EED (embryonic ectoderm development, Esc for Extra sex combs in *Drosophila*), SUZ12 (Su(z)12 for Supressor of zeste 12 in *Drosophila*) and RBBP4/7 (for retinoblastoma binding protein 4/7, also known as RbAp48 or NURF55, and RbAp46). Eed and Suz12 activate the catalytic activity of EZH2 (Cao and Zhang, 2004; Pasini et al., 2004) by inducing conformational changes in EZH2, which is otherwise auto-inhibited (Antonysamy et al., 2013; Wu et al., 2013a). After H3K27me3 mark has been deposited by EZH2, it can be recognized by EED through its WD40 domain, which in turn not only brings the PRC2 complex to the marked site, but also further allosterically activates the enzymatic activity of EZH2 to modify the neighbouring nucleosomes (Jiao and Liu, 2015) which causes spreading of this mark to nearby nucleosomes (Margueron et al., 2009; Oksuz et al., 2018;

Poepsel et al., 2018). Of note, this allosteric activation is minimal in the case of EZH1-containing PRC2 complexes, and they essentailly can only deposit H3K27me1 and H3K27me2 marks (Lee et al., 2018). Of note, the H3K27me2 mark is found on >70% of H3 molecules in ES cells (Jung et al., 2010) and essentially blankets the whole genome with exception of regions with H3K27me3, and partially depleted over actively transcribed regions (Ferrari et al., 2014). Interestingly, this correlates with the recent finding of pervasive low-level H2AK119ub mark deposited by PRC1 (Fursova et al., 2019), but whether these two observations are linked is unkown.

The RBBP proteins are required for nucleosome binding by the PRC2 complex (Nekrasov et al., 2005), and also stimulate its enzymatic activity (Cao and Zhang, 2004).

While spreading via EED binding of H3K27me3 sustains high level of PRC2 binding at its target sites, it is not sufficient to maintain binding and repression during cell divisions when modified nucleosomes get rapidly diluted: continuous sequence-specific targeting is required to avoid this (in Drosophila, where this was shown, PRC2 binds specific sequences termed Polycomb Response Elements, PREs)) (Coleman and Struhl, 2017; Laprell et al., 2017). Moreover, re-expression of PRC2 components in knock-out cells that lost all H3K27me3 restores normal patterns of H3K27me3, which can’t be explained by local propagation of PRC2 recruitment based on the H3K27me3 mark (Højfeldt et al., 2018). The mechanisms of PRC2 targeting to specific sites in mammalian cells are still under active investigation and, similarly to PRC1, has been challenging to characterize (Yu et al., 2019). Two of the many proposed models

are based on either sequence-specific, or chromatin-mediated recruitment. Core components of PRC2 do not contain DNA-binding or any additional nucleosome-binding activities, so other accessory factors are required for this. The main candidate proteins are JARID2 and MTF2. MTF2 (Metal Regulatory Transcription Factor 2, also known as PCL2, for Polycomb-Like Protein 2) and other PCL proteins are able to recruit PRC2 to CpG islands by binding to a specific motif and/or DNA shape (Casanova et al., 2011; Li et al., 2017; Perino et al., 2018). JARID2 (Jumonji and AT Rich Interactive Domain 2), on the other hand, can target PRC2 to H2AK119ub modified chromatin (Cooper et al., 2016). Additionally, PCL proteins, JARID2 and other accessory components of PRC2 increase the general affinity of the complex to chromatin, and regulate PRC2 activity (Lee et al., 2018; Son et al., 2013; Wang et al., 2017a).

While I am not going to discuss it in detail, a lot of studies recently have focused on interaction between PRC2 and RNA, and it has been suggested both as a targeting mechanism, and a way to mask the genes from silencing by PRC2 (reviewed in Yan et al., 2019).

1.3.3 Models of PRC Recruitment

For a long time the dominant model of PRC1/2 recruitment was the hierarchical model (Cao et al., 2002; Wang et al., 2004b). It includes binding of PRC2 first, deposition of H3K27me3 mark by EZH2, which is then recognized by EED to promote domain maintenance and spreading, and by CBX proteins to recruit PRC1. More recently a number of observations have questioned the validity, or at least the completeness of this simple model. First, it was found that during X chromosome inactivation, which normally involves both PRC1 and PRC2

binding through their interaction with the long non-coding RNA XIST, PRC1 is recruited normally in cells lacking a core component of PRC2, EED (Schoeftner et al., 2006). Later it was shown that vPRC1 is recruited to its normal targets in mouse ES cells independently of PRC2 (Tavares et al., 2012). A mechanism for this was proposed: KDM2B, a component of PRC1.1, can bind CGIs, and bring PRC1 to them (Farcas et al., 2012; He et al., 2013; Wu et al., 2013b); this, however, didn't explain recruitment of PRC2 to the same sites. This was later proposed to occur via H2AK119ub mark, deposited by vPRC1 and recognized by PRC2 (Blackledge et al., 2014; Cooper et al., 2014; Kalb et al., 2014). Specifically, JARID2 has been implicated in recognizing H2AK119 mark (Cooper et al., 2016).

At this point, however, conflicting results were reported. The I53A mutation of RING1B disrupts its interaction with the E2 ubiquitin ligase and therefore prevents deposition of H2AK119ub mark (Buchwald et al., 2006; Elderkin et al., 2007). The RING domain which performs the ubiquitin ligase function is also important for dimerization of RING1B with the PCGF component of PRC1. The other reported mutation R70C that renders RING1B catalytically inactive (Wang et al., 2004a), likely disrupted this interaction since this residue is directly involved in the salt bridge formation with PCGFs (Buchwald et al., 2006), while the I53A mutations has been shown to allow the correct complex formation (Buchwald et al., 2006; Elderkin et al., 2007). I53A mutation turned out to have a surprisingly mild effect on gene repression in mouse ES cells, development of mouse embryos and skin (Cohen et al., 2018; Illingworth et al., 2015). Introduction of the equivalent I48A mutation into the *Drosophila* Sce E3

ligase resulted in flies many of who didn't develop into adults, but those that did had no homeotic transformations characteristic of PRC1/2 mutants, although some variable more subtle phenotypes were observed (Pengelly et al., 2015). Of note, these flies contained maternally deposited wild-type Sce which remained active in until late embryo stages. In the same study, authors generate flies with mutations of the modified H2A residues to prevent any possibility of H2A ubiquitination, and these animals arrested development at the end of embryogenesis, although they, again, contained maternally deposited wild-type histones. However, the authors analysed clones of cells containing only non-modifiable histones (including H2Av, H2A.Z homologue) in imaginal discs, and showed no de-repression of HOX genes in this condition.

Finally, recently two papers add more details to this puzzle. In one of them a series of conditional PCGF knock-out ES cells was created, and analysis of these suggest that vPRC1 complexes are key for gene regulation, and their removal not only abrogates gene repression, but also causes loss of H2AK119ub, while cPRC1 complexes are mostly dispensable in ES cells (Fursova et al., 2019). Interestingly, a subset of PRC1 targets maintained some level of repression in PCGF1/3/5/6 knockout cells relative to RING1B knock-out cells, and these were particularly extended with very high levels of PRC1 binding in wild type, which included, for example, *Hox* genes, and these are probably silenced by cPRC1 at least partially independently of vPRC1. Also, *in vivo* mutations of canonical components cause severe developmental phenotypes in mice (Isono et al., 2005; Lau et al., 2017), and therefore cPRC1 is important for correct development. The other paper investigates in detail the

role of catalytic activity of PRC1 by showing that RING1B I53A is not completely enzymatically inactive, but is severely hypomorphic with very low residual activity (Blackledge et al., 2019). They then generate cells with a double point mutation I53A/D56K that fully abrogates the RING1B catalytic activity *in vitro*, and show that this causes complete loss of H2AK119ub. This leads to a loss of binding of PRC2 components and H3K27me3, while I53A cells contain low level residual modification, probably for correct recruitment of PRC2. Similarly to the previous study, a small subset of genes maintained some level of repression in cells with catalytically dead RING1B, and they were also particularly long regions with high level of cPRC1 binding. The above suggests that there are two subclasses of polycomb targets that have differential reliance on cPRC1 and vPRC1 activity: a small subset of canonical targets, including *Hox* genes, bound to a large extent by cPRC1, and the rest that relies essentially only on vPRC1.

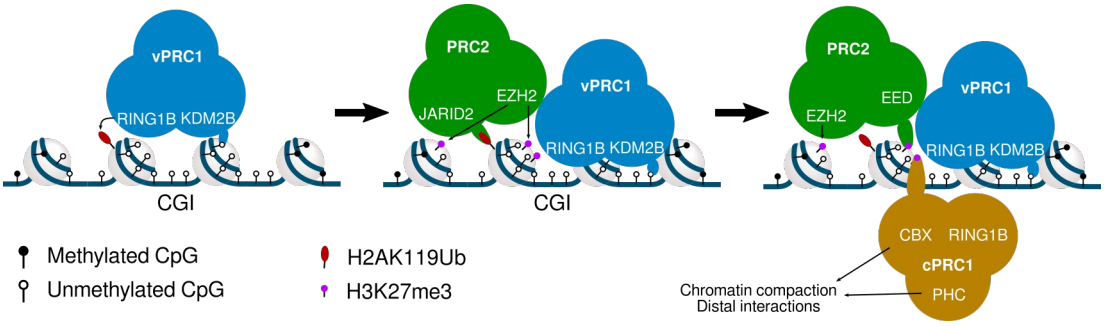

Figure 1.10. Simplified schematic of PRC1/2 recruitment. It starts by recognition of unmethylated DNA by vPRC1 and deposition of H2AK119ub. This mark is then recognized by PRC2 (probably though the JARID2 component), which then deposits H3K27me3. This mark is in turn recognized by cPRC1 through CBX subunits, and by PRC2 itself.

The overall simplified model of vPRC1-PRC2-cPRC1 recruitment is presented in Figure 1.10. The original hierarchical model (Cao et al., 2002; Wang et al., 2004b) only includes the second and third stages with PRC2 recognizing its targets without use of H2AK119ub. While often believed to be outdated (Blackledge et al., 2014), it is worth mentioning that PRC2 recruitment happens independently of PRC1 on a subset of loci: some CGI promoters are covered by H3K27me3 and bound by PRC2 components, but have no detectable PRC1 binding (Ku et al., 2008).

1.3.4 Polycomb bodies

It has long been observed that various PcG proteins form discreet foci in the nucleus, both in *Drosophila*, and in mammalian cells (Buchenau et al., 1998; Saurin et al., 1998). These were termed “Polycomb bodies”, and have been thought to be one of the mechanisms of PcG-mediated gene repression. A number of proteins have been implicated in their formation, however the prevalent model is that they are formed by cPRC1 complexes (reviewed in Illingworth, 2019).

PHC was shown to be involved in formation of these structures via its SAM domain oligomerization, that drives local compaction of PRC1 targets and/or brings distal targets into proximity (Isono et al., 2013). Interestingly, a single point mutation in the SAM domain of overexpressed PHC2 disrupted RING1B clusters in wild-type human cells, indicating a dominant effect of the non-oligomerizing protein. This mutation also caused loss of compaction of the HoxB gene cluster in MEFs and derepression of some of the genes, and homozygous mice displayed skeletal phenotypes characteristic of Hox gene

misregulation. While suggestive of the role of Polycomb bodies in gene regulation, the authors also find a severe loss of RING1B binding and H3K27me3 mark at promoters, therefore the derepression could be caused by lower occupancy of PRCs, not due to lack of polycomb body formation; similarly, whether Polycomb bodies are lost due to the direct effect of the PHC mutation or due to derepression of target genes is also unclear.

In *Drosophila*, deletion of a Polycomb response element within one of the Hox clusters causes partial derepression of the other Hox cluster, and a developmental phenotype (Bantignies et al., 2003, 2011). This suggests that clustering of distal PRC1 targets plays a role in their silencing. However a similar experiment has not been performed in mammalian cells.

The CBX proteins, and in particular CBX2, has also received a lot of attention in the context of Polycomb body formation. First, CBX2 (also known as M33) has been shown to compact nucleosomal arrays *in vitro* and inhibit remodelling by SWI/SNF complexes, and this was suggested as one of the mechanisms for compaction and silencing of PRC1 targets (Grau et al., 2011). A stretch of positively charged residues was required for this activity, and later was shown to be important for PcG-mediated silencing *in vivo* in mice (Lau et al., 2017). Recently it was proposed that CBX2 protein can undergo phase separation (see 1.1.2.4), and this requires the presence of positively charged residues in the protein (Plys et al., 2019; Tatavosian et al., 2018). Phase separation properties could mediate the formation of Polycomb bodies and/or local compaction of Polycomb targets.

A few studies have addressed the question of which regions form Polycomb bodies in mammalian cells, and are locally compacted and/or take part in distal interactions. One of the first papers addressing this focused on the Hox gene clusters, and analysed their compaction state by FISH: measuring the distance between probes at either end of the Hox cluster as a proxy for compaction (Eskeland et al., 2010). Both HoxB and HoxD regions were highly compacted in ES cells, but decompaction was observed upon Retinoic acid differentiation and concomitant marked loss of H3K27me3 at these regions. Similarly, loss of compaction was observed in EED and RING1B knock-out cells, while Hox cluster organization was not altered in RING1B I53A cells. Since according to the hierarchical model of PRC1/2 recruitment, prevalent at the time, in EED KO cells PRC1 binding would be lost, while loss of RING1B would not affect binding of PRC2, the authors concluded that PRC1, and not PRC2 is the likely driver of local compaction.

With the advent of C-methods, several studies have addressed the role of PRCs in spatial genome organization, often using genome-wide approaches, in mammalian cells. Using Promoter-Capture Hi-C, interaction network of PRC1 targets in ES cells was analysed (Schoenfelder et al., 2015). It reported that PRC1/2 binding sites are frequently found interacting with each other, and this depends on RING1A/B. Interestingly, this analysis suggests a special role for Hox loci which form a “Hox network” with particularly high enrichment of interactions: both between Hox clusters, between Hox clusters and some other Polycomb targets, and between the other Polycomb targets in the network.

However, no evidence was presented to show any special properties of Hox regions beyond their size and level of PRC1/2 binding.

Another publication performed Capture Hi-C on DNase I hypersensitive sites in ES cells, grown in serum and 2i (Joshi et al., 2015), after the same group showed reorganization of PRC2 binding in 2i culture and its extensive loss at CGIs (Marks et al., 2012). Here they describe Extremely Long-Range Interactions (ELRIs) between H3K27me3-marked sites. These interactions were lost or markedly reduced in ground state ES cells, consistently with loss of PcG binding in this condition, and often occurred over very long distances (10s of Mbp). They also performed this analysis in EED knock-out ES cells and observed complete loss of ELRIs. They went on to show that sites that form ELRIs have particularly high levels of SUZ12 and RING1B binding. Their analysis did not however distinguish which PRC complexes were directly involved in mediating these interactions.

In the same year another publication reported very similar findings regarding long-range looping between PRC1/2 targets in wild-type cells using 4C and reanalysis of Hi-C data (Vieux-Rochas et al., 2015). They found interactors of Hox regions from 4C data, which turned out to be high-occupancy H3K27me3 regions. They also showed that these regions can interact across chromosomes, are found in the active (A) compartment according to Hi-C data, and are not in LADs. However here, again, the analysis didn't identify, which PRC complexes were responsible for looping interactions. Like the previously discussed report, while uncovering the presence of long-range contacts, these results cannot be used to quantify whether the long-range interactions occur

less frequently than short-range interactions, relative to expected level of interactions from just genomic background.

A different approach was taken in another recent report (Kundu et al., 2017): using a combination of super-resolution FISH imaging and high-resolution 5C (a 3C-based method for interrogation of a specific genomic region), authors investigate local compaction and short-range looping in a selection of regions containing major PRC targets, including Hox clusters and other regions. They show loss of compaction in ES cells upon loss of PHC1, implicating cPRC1 in performing local compaction. Similar results were observed when ES cells were differentiated into Neural Progenitor Cells (NPCs) accompanied by loss of PRC1/2 binding. In the 5C data, loss of contacts between neighbouring PRC1/2 targets, such as the Nkx2-2, Nkx2-4 and Pax1 genes, was also observed. The authors went on to reproduce results from (Eskeland et al., 2010) using 5C to show loss of compaction and looping in RING1B knock-out cells, but their preservation in RING1B I53A mutant ES cells. While performed only on a small set of regions, this study is the most direct evidence that cPRC1 is the subtype of PRC1 that is key for creating local compaction and interaction between PRC targets.

A study that used high resolution Hi-C addressed the dynamics of PcG-associated interactions during neuronal differentiation (Bonev et al., 2017). These interactions were observed to occur both within TADs, between TADs, and at long distance separations. Level of interactions in Hi-C correlated much better with RING1B than H3K27me3, consistently with the role of PRC1 in driving looping. Interestingly, they observe loss of these loops during

differentiation, related to the decrease in level of RING1B binding with constant H3K27me3. Strikingly, loss one of the interactions between HoxA gene cluster and the Tlx2 gene across ~30 Mbp during the differentiation could be confirmed by FISH.

Most recently, two studies addressed the question whether compaction and/or looping between PRC1/2 targets are related to cohesin and loop extrusion. First, Rhodes and colleagues developed RING1B-AID and SCC-AID mouse ES cells, and showed that upon acute depletion of RING1B interactions between PRC1/2 targets are lost, however depletion of SCC1 (and therefore the cohesin complex) makes RING1B-associated interactions stronger (Rhodes et al., 2020)(Rhodes et al., 2020)(Rhodes et al., 2020), consistently with loop extrusion activity causing mixing of compartments and preventing interactions between super-enhancers (Mirny et al., 2019; Rao et al., 2017a; Schwarzer et al., 2017). The second study investigated the roles of the SA1 and SA2 subunits of cohesin by siRNA mediated depletion (Cuadrado et al., 2019). Their analysis suggests that SA1 antagonizes RING1B-associated interactions, while SA2 is needed for their formation. Together these studies suggest that either SA2 has a role outside the canonical SMC1/SMC3/SCC1 cohesin complex, related to Polycomb, or that cohesin-SA2 is required for maintenance of PRC1 domains and 6 hours-long depletion of SCC1 in the previous study was not sufficient to cause loss of PRC1 binding, while the longer siRNA depletion of SA2 allows for PRC1 binding to decrease.

1.3.5 Interplay of polycomb with DNA methylation

As mentioned above, PRCs primarily bind CGIs. This can be achieved through KDM2B targeting PRC1.1 to CGIs (Farcas et al., 2012; He et al., 2013; Wu et al., 2013b), or JARID2 or Polycomb-like proteins (PCL, including MTF2) targeting PRC2 (Casanova et al., 2011; Li et al., 2010, 2017). It is unclear how the CGIs are chosen, since the mentioned above proteins have affinity to all CpG-rich DNA. It is possible that PRC2 is restricted from binding some CGIs due to presence of certain activating signals, such as RNA (Cifuentes-Rojas et al., 2014; Kaneko et al., 2014) or H3K4 and H3K36 methylation (Schmitges et al., 2011).

The generic affinity of PRCs to unmethylated CpG-containing DNA, and inhibitory effects of DNA methylation on PRC binding, creates the possibility of affecting PRC distribution by affecting DNA methylation patterns. Loss of DNA methylation would allow binding of PRCs to a vastly bigger repertoire of CpGs outside of their normal target regions. A key paper reported redistribution of PRC2 binding in mouse embryonic fibroblasts (MEFs) lacking *Dnmt1*, the maintenance DNA methyltransferase. Regular PRC2 target CGIs lost the bulk of H3K27me3 mark, which caused derepression of target genes (Reddington et al., 2013). Similarly, a loss of H3K27me3 over CGIs was reported in ES cells lacking all three DNMTs (TKO cells) (Brinkman et al., 2012). In another study, relocation of PRCs to DAPI-dense chromocentres containing major satellites in ES cells was observed in DNA hypomethylation conditions, either in TKOs or in *Uhrf1*$^{-/-}$ cells (Cooper et al., 2014). Interestingly, in cancers, CGIs that gain methylation in tumours were found to be silenced by Polycomb in the tissue of

origin (Reddington et al., 2014), consistently with the overall reciprocal relationship between these two repressive states.

Culture of wild-type ES cells in 2i media that promotes ground state pluripotency (see 1.4.2 for details) causes global hypomethylation of the genome, and induces redistribution of PRC binding (Joshi et al., 2015; Marks et al., 2012) away from CGIs.

This trend of redistribution of PRC binding during changes in DNA methylation is biologically very interesting due to the global changes in DNA methylation levels during development (see 1.2.4) (reviewed in Lee et al., 2014). This suggests the distribution of PcG binding across the genome might be highly dynamic due to global changes in the DNA methylone during development, however this question has not been addressed in detail in the literature.

1.4 Mouse embryonic stem cells

Mouse embryonic stem (ES) cells (or mESCs) are a unique cell type, cultured *in vitro* obtained from cells present in the early embryo *in vivo* (Evans and Kaufman, 1981; Martin, 1981), which retains some of the key properties required to support the development of the organism: they can be differentiated into any other embryonic cell type by simply changing the culture conditions, and they self-renew, i.e. effectively indefinitely proliferate without losing the differentiation potential. These cells, when injected into the blastocyst, can generate chimeric mice with adult cells coming from both the host blastocyst and from injected ES cells (Bradley et al., 1984). Since this discovery, this property of ES cells has been widely used for generation of mutant mouse lines: genetic manipulation in ES cells, amenable to normal cell culture methods, allows generation of mice carrying the altered genome (Robertson et al., 1986; Thompson et al., 1989).

1.4.1 Growth conditions of mouse ES cells

ES cells were originally derived by culturing the E3.5 blastocyst (or the inner cell mass (ICM)) on a layer of feeder cells – mitotically inactivated fibroblasts in the presence of foetal calf serum (FCS, or simply serum) (Evans and Kaufman, 1981). Later it was discovered that the main factor provided by the feeder cells is the cytokine Leukemia Inhibitory Factor (LIF) (Smith et al., 1988), and growth of ES cells feeder-free with supplemented LIF is now common practice. Mechanistically, LIF influences gene expression via the JAK/STAT pathway, and activates the components of the pluripotency network, such as *Klf4*, *Gbx2*

and *c-Myc* (Cartwright et al., 2005; Hall et al., 2009; Niwa et al., 2009; Tai and Ying, 2013).

Understanding what components of FCS are required for culture of ES cells was more challenging, but Bone Morphogenetic Protein (BMP) turned out to be the key component of serum (Ying et al., 2003). It activates *Id* (*Inhibitor of Differentiation*) genes and E-Cadherin expression via SMAD signalling (*small* and *Mothers Against Decapentaplegic)* (Malaguti et al., 2013), and suppresses ERK and p38 (Qi et al., 2004). LIF and BMP together suppress ES cell differentiation. Omitting either of these factors from the growth medium causes spontaneous differentiation, and BMP on its own even promotes differentiation into non-neural tissues (Koopman and Cotton, 1984; Smith and Hooper, 1983; Wiles and Johansson, 1999). Therefore ES cells have an intrinsic differentiation cue, that gets suppressed by culture with these signalling molecules. These signals turned out to be the FGF4 (fibroblast growth factor 4) signalling, that acts through the ERK (Burdon et al., 1999; Wilder et al., 1997), and *Erk* mutant ES cells don't require the presence of BMP in the culture medium to avoid spontaneous differentiation (Kunath et al., 2007). Similarly, chemically inhibiting the Mek1/2 kinases required to stimulate ERK together with FGF4 receptor inhibition allowed culture in serum-free conditions (Ying et al., 2008). GSK3 (glycogen synthase kinase-3) was another protein shown to be antagonistic to self-renewal of ES cells (Doble et al., 2007; Sato et al., 2004) through inhibition of β-catenin signalling. These observation combined with availability of selective small molecule inhibitors of Mek1/2 and GSK3 (PD0325901 and CHIR99021, respectively) allowed development of serum-free medium for ES

cell culture termed “2i” (for **2 i**nhibitors) (Silva et al., 2008; Ying et al., 2008). Devoid of differentiation cues present in serum (such as BMP), cells cultured in 2i enter “ground state pluripotency” - instead of variable levels of pluripotency factors between cells in the population due to temporary priming towards different lineages (Canham et al., 2010; Hayashi et al., 2008), they display uniform levels of such key pluripotency factors like *Oct4*, *Nanog* and *Prdm14* (Yamaji et al., 2013) (Figure 1.11). Therefore, cells cultured in 2i are more “naïve” and homogenous than the serum-cultured cells that correspond to slightly later stages in development, although this priming is reversible and generates a meta-stable pluripotent population (Abranches et al., 2013; Leitch et al., 2013).

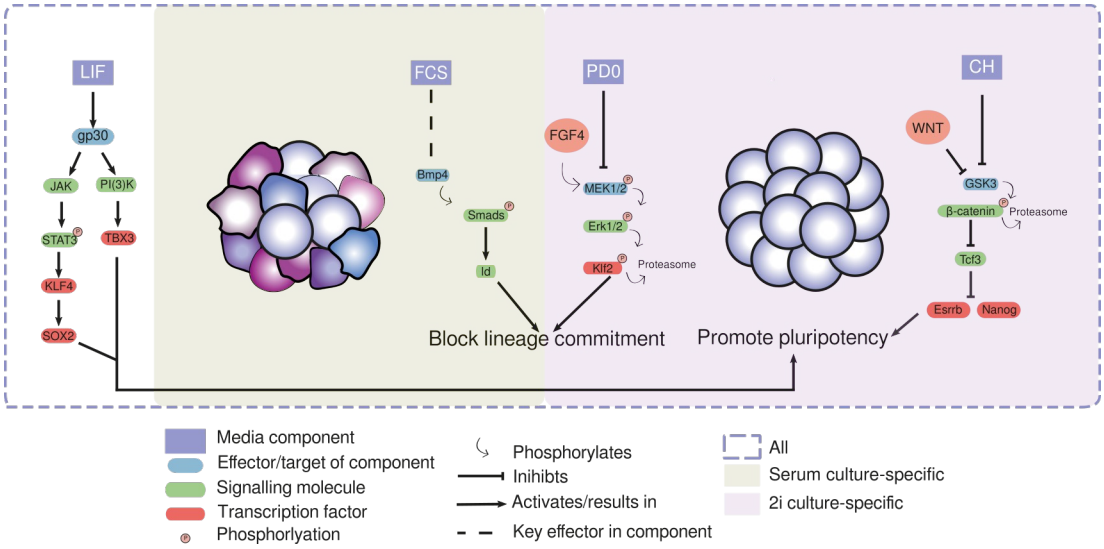

Figure 1.11. Key pluripotency signalling pathways in serum and 2i mouse ES cells culture conditions. Adapted from figure by Katy McLaughlin (McLaughlin, 2018).

1.4.2 Epigenome of mouse ES cells in serum and 2i culture

The epigenome of mouse ES cells undergoes a major reorganization between serum and 2i culture. In particular, DNA methylation levels and PRC1/2 binding patterns are particularly affected.

1.4.2.1 Loss of DNA methylation in 2i culture

DNA methylation is globally greatly reduced in ground state ES cells (Ficz et al., 2013; Habibi et al., 2013; Leitch et al., 2013). Interestingly, the level of 5mC in 2i cultured cells closely matches that of ICM, while serum-cultured ES cells resemble the E.6.5 epiblast cells (Figure 1.8) (Ficz et al., 2013). Importantly, however, genome-wide the patterns of methylation of ES cell genomes are distinct from *in vivo* observations at any stage, and resemble more the populations in the early postimplantation embryos, suggesting even 2i cells, while recapitulating the levels, don't model the epigenome of ICM cells (Zhang et al., 2018).

Loss of DNA methylation in the ground state is observed genome-wide with only few regions retaining levels from serum-cultured ES cells: imprinted regions and some IAP elements are the only loci fully protected from demethylation (Ficz et al., 2013; Habibi et al., 2013). Only for a small subset of loci, loss of DNA methylation might be associated with activation of gene expression, since the same genes are activated in *Dnmt1/3a/3b* triple knockout ES cells with no DNA methylation (Leitch et al., 2013). However the genes with known methylation-sensitive promoters, such as *Dazl* and *Gstp2*, are

upregulated with loss of methylation. While the majority of repetitive elements lose DNA methylation, their expression is not increased.

The loss of DNA methylation in 2i culture is linked to reduced expression of *Dnmt3a/3b/3l* therefore *de novo* DNA methylation activity is impaired (Ficz et al., 2013; Habibi et al., 2013; Leitch et al., 2013). Moreover, another report suggests that maintenance methylation is also perturbed due to loss of UHRF1 protein and its association with replication forks (von Meyenn et al., 2016). It has also been suggested that TET enzymes might be facilitating DNA demethylation during serum-2i transition at some loci (Ficz et al., 2013; Hackett et al., 2013).

PRDM14 (PR Domain 14) is a key transcription factor required for hypomethylation in 2i conditions (Ficz et al., 2013; Grabole et al., 2013; Habibi et al., 2013; Leitch et al., 2013; Yamaji et al., 2013). PRDM14 prevents differentiation, and moreover its depletion causes upregulation of *Dnmt3b* and *Dnmt3l* and increased level of DNA methylation (Ma et al., 2011; Yamaji et al., 2013). Interestingly, PRDM14 is upregulated in 2i, which correlates with loss of DNA methylation and *de novo* DNMT expression (Ficz et al., 2013; Hackett et al., 2013), and it's induced expression leads to decreased *Dnmt3a*, *Dnmt3b*, and *Dnmt3l* expression (Hackett et al., 2013). PRDM14 has been shown to be a sequence-specific TF (Ma et al., 2011), and to co-bind genomic sites with ESRRB, NANOG and other key pluripotency regulators (Ma et al., 2011; Yamaji et al., 2013). It has also been suggested to silence gene expression by recruitment of PRC2, which might be how it regulates *de novo* DNMTs (Yamaji et al., 2013).

1.4.2.2 Redistribution of Polycomb binding in 2i

Unlike other investigated histone marks (H3K4me3, H3K36me3 and H3K9me3), H3K27me3 distribution is markedly different between primed and ground state ES cells (Marks et al., 2012). In particular, it is reduced at its normal targets (CGI promoters), but gained at some repeat elements, such as satellites, with overall similar levels of H3K27me3 globally. Interestingly, loss of H3K27me3 at promoters was not associated with loss of repression despite the repressive functions of PRC complexes. Binding of EZH2, SUZ12 and RING1B is similarly perturbed (Joshi et al., 2015; Marks et al., 2012). This is consistent with the intrinsic affinity of PRC1/2 complexes towards unmethylated DNA, and their titration away from CGIs upon global genome demethylation (Brinkman et al., 2012; Jermann et al., 2014; Reddington et al., 2013).

While the effect this redistribution of PRC1/2 binding has on 3D organization of the genome has not been studied extensively and is one of the objectives of this work, one report investigated the changes in looping between DNase I hypersensitive sites using capture Hi-C, and reported extensive loss of interactions between Polycomb targets, in particular at long distances (Joshi et al., 2015). Another recent study used Hi-C to globally analyse 3D genome organization in ES cells, and reported depletion of RING1B-associated interactions in 2i in *cis* and in *trans* (Cuadrado et al., 2019), however their analysis is not comprehensive.

1.5 Aims of this study

In my PhD I intended to investigate the 3D organisation of the genome of mouse ES cells driven by PRC1 genome-wide using Hi-C and computational analysis.I first developed a new tool for versatile Hi-C analysis to create pile-up plots used for quantification of interactions between genomic regions. I then applied it, together with other analysis methods, to Hi-C data from wild-type, RING1B$^{I53A/I53A}$ and RING1B$^{-/-}$ mouse ES cells to investigate PRC1-mediated local compaction and looping, and their interplay with local features of PRC1 binding sites. I then took advantage of the 2i culture system as a model of the natural developmental process with loss of PRC1 binding from its usual targets due to DNA demethylation to investigate the effect it has on 3D genome organization.

Chapter 2: Materials & Methods

1.1 Cell culture

mESCs used in relation to Chapter 4, were cultured by Rob Illingworth. They were maintained at 37° with 5% CO_2 feeder-free on 0.1% gelatin (Sigma G1890) coated Corning flasks in GMEM BHK-21 (Gibco 21710-025). The medium was supplemented with 10% foetal calf serum (FCS Sigma F-7524), 1000 units/ml Leukemia inhibitory factor (LIF; produced in-house), 2mM L-Glutamine (produced in-house), 1 mM Sodium Pyruvate (Gibco 11360-039), 50 mM 2-β-mercaptoethanol (Gibco 31350-010). For passaging, 60-90% confluent flasks were washed with PBS, treated with trypsin (0.05% v/v; Gibco 25300-054) for 2-3 minutes (mins) at room temperature (RT), and tapped to detach the cells. Nine volumes of complete medium were used to inactivate the trypsin, the mixture was repeatedly pipetted to prepare single-cell suspension. After centrifugation (300 g), mESCs were counted using a haemocytometer and plated onto a new flask at a density of approximately 4×10^4 cells/cm^2.

mESCs used in relation to Chapter 5, were cultured by Katy McLaughlin. They were maintained at 37° with 5% CO_2 feeder-free on 0.2% gelatin (Sigma G1890) coated Corning flasks. Serum cultured cells were grown in GMEM (Gibco) supplemented with 15% foetal calf serum, 0.1 mM nonessential amino acids (Sigma), 1 mM Sodium Pyruvate (Sigma) 1% Penicillin/Streptomycin, 2 mM L-glutamine, 0.1 mM β-mercaptoethanol (Thermo Fisher), and ESGRO LIF (Millipore) at 1000 U/mL. 2I conversion was performed over 14 days by growing cells in medium including 50% DMEM/F12 (Gibco), 50% Neurobasal media (Gibco), 0.5% N2 supplement, 1% B27 & RA (Gibco), 7.5% BSA (Gibco), 1% Penicilllin/Streptomycin, 2 mM L-glutamine, 0.15 mM

monothioglycerol (Sigma), 1000 U/ml ESGRO LIF (Millipore), 1 μM PD0325901 (MEK inhibitor, Stemgent) and 3 μM CHIR99021 (GSK3 inhibitor, Stemgent).

1.2 *In situ* Hi-C

Hi-C analysis was performed as described (Rao et al., 2014) with minor modifications.

After resuspending mESCs from a near-confluent T75 flask like for passaging, they were counted (approximately 10 million cells were typically obtained), washed once in 10 ml Phosphate-buffered saline (PBS; produced in house) and centrifuged at 300g for 5 min. For crosslinking, they were resuspended in DMEM or GMEM with 1% formaldehyde (Thermo Scientific 28908, or CALBIOCHEM 344198); approximately 1 ml of fixing solution for each million of cells), and were incubated for 10 mins on a rocker. Formaldehyde was quenched by addition of 2 M glycine solution to a final concentration of 0.2 M, then the cells were incubated for 5 min at RT on a rocker. Cells were centrifuged for 5 minutes at 300g. The supernatant was discarded. The cells were resuspended in 1 ml of PBS and centrifuged at 2500g for 5 min. Then they were again resuspended in 1 ml PBS and split into aliquots of 2-5 million cells. All aliquots were centrifuged again at 2500g for 2 min, and supernatant was discarded. If the cells were to be processed further at a later time, the tubes with cell pellets were snap-frozen in liquid nitrogen and stored at -80°C for months. The pellet was then resuspended in 300 µl Hi-C lysis buffer (10 mM Tris-HCl pH 8.0, 10 mM NaCl, 0.2% IGEPALCA-630 (SIGMA I8896), 1x protease inhibitor Thermo Scientific, 78430). The cells were incubated on ice for 15 min, then centrifuged for 5 min at 2500g at 4°C. Supernatant was discarded, the pellet was resuspended in 500 µl lysis buffer, and centrifuged again. After discarding the supernatant, the cells were further lysed in 50 µl

0.3% sodium dodecil sulphate (SDS) in 1×NEBuffer 3 (New England Biolabs) at 62°C for exactly 10 minutes. SDS was inactivated by addition of 147.5 µl NEBuffer 3 and 12.5 µl Triton X-100 (Sigma 93443), careful mixing and incubation at 37°C for 1 hour with shaking. After this, a 30 µl chromatin integrity control sample was taken and frozen at -20°C. The nuclei were centrifuged at 3000g for 5 min and then resuspended in 250 µl of 1×DpnII buffer with 600 units of DpnII restriction enzyme (NEB) to digest the chromatin at 37°C overnight with shaking. In the morning 200 units of DpnII were added for extra digestion for 2 hours.

After the chromatin was fragmented by DpnII, the enzyme was inactivated by heating the samples at 65°C for 20 mins. A 40 µl digestion control sample was then taken and frozen at -20°C. 40 µl mQ water was added to the samples to replenish the volume. The ends of the digested DNA were filled in to incorporate biotin and mark future ligation junctions by addition of 50 µl of a fill-in master mix (0.3 mM of each dCTP, dGTP, dTTP (Life Technologies 10297018) and biotin-14-dATP (Invitrogen 19524-016), and 40 Units of DNA Polymerase I Klenow Fragment (NEB M0210)). Samples were carefully mixed and incubated at 37°C for 1.5 hrs with shaking. Blunted ends of DNA were ligated *in situ* by adding 900 µl of ligation master mix (1.33× NEB T4 DNA ligase buffer, 1.11% Triton X-100, 1.33× BSA (NEB, or in house prepared according to the NEB specification), 2000 cohesive end units T4 DNA ligase (NEB M0202M) and incubating for 4 hours at RT with rotation. The nuclei were centrifuged at 3000g for 10 mins, then resuspended in 200 µl mQ water. Protein was degraded by addition of 10 µl 10 or 20 mg/ml proteinase K and 10

µl 20% SDS and incubation at 55°C for 30 mins. The same was done with earlier collected control samples after adjusting their volumes to 200 µl. Then, 130 µl of 5M NaCl was added to the tubes and they were incubated at 65°C overnight.

In the morning, the tubes were cooled to RT. Then, 610 µl of absolute ethanol (EtOH) and 38 µl of 3M sodium acetate were added, samples were mixed and incubated at -80°C for 15 mins. DNA was precipitated by centrifugation at top speed in a table-top microcentrifuge at 4°C for 15 mins. The supernatant was carefully removed and the pellets were either washed with 500 µl of 70-80% cold EtOH by centrifuging the tubes for 5 mins and resuspended in 30 µl 10 mM Tris-HCl pH 8.0, or resuspended in 500 µl of 10 mM Tris-HCl pH 8.0 and washed on Amicon filter units (30K 500 µl UFC5030BK) twice. 5 µl of each sample was taken as ligation control and added to 20 µl10 mM Tris-HCl pH 8.0. Sonication buffer (50 mM Tris-Hcl pH 8.0, 0.1% SDS, 10 mM EDTA) was added to the samples to total volume of 500 µl, and they were let stand on ice for 15 mins. Afterwards the DNA was fragmented using a probe sonicator to achieve a fragment size distribution of ~200-700 bp. Fragmented DNA was washed twice on Amicon Filter units, and then eluted. One µl of the samples was transferred to Qubit assay tubes for measurement of DNA concentration, together with control samples. Two hundred ng of control samples were run on 1% agarose/TBE gel to check digestion and ligation efficiency.

To bind biotinylated DNA to streptavidin beads (Dynabead MyOne Streptavidin T1, Life Technologies 65602), 30 µl of the beads per sample were washed with 80 µl of Tween Wash buffer (TWB: 5mM Tris-HCl pH 7.5, 0.5mM EDTA, 1M

NaCl, 0.05% Tween 20) per sample. The supernatant was removed after magnetic separation of the beads from the solution. The beads were resuspended in 50 µl of 2× binding buffer (BB: 10mM Tris-HCl pH 7.5, 1mM EDTA, 2M NaCl) per sample. Then an equal volume of 2×BB was added to each sample in DNA LoBind tubes (Ependorf 0030108051) and they were incubated for 15 mins at RT with rotation. The supernatant after bead separation was kept in a separate tube. The beads were washed twice with TWB at 55°C for 2 mins with shaking, and transferring into a new tube each time. The supernatant from the first wash was combined with the previously collected supernatant. This solution was used to quantify DNA concentration to estimate biotin binding efficiency; good libraries had ~40-60% binding efficiency. The beads were washed with 1× T4 DNA ligase buffer. To repair the ends and remove biotin from unligated ends the beads were resuspended in 100 µl end-repair mix (85 µl 1x T4 DNA ligase buffer (NEB), 5 µl 10 mM dNTPs, 5 µl T4 Polynucleotide kinase (NEB M0201L), 4 µl T4 DNA polymerase (NEB M0203L), 1 µl DNA polymerase I Klenow fragment (NEB M0210) and incubating the samples at RT for 30 min. Afterwards the beads were washed twice with TWB, and then once in 100 µl of 1× NEBuffer 2. Then the repaired ends were A-tailed by resuspending the beads in 90 µl 1× NEBuffer 2, 5µl 10 mM dATP and 5 µl DNA polymerase I Klenow (3`→5` exo-) fragment (NEB, M0212L) and incubation at 37°C with shaking for 30 mins. The beads were again washed twice with TWB, and then once with 100 µl T4 DNA ligase buffer. Then Illumina sequencing adaptors were ligated by resuspending the beads in 7 µl mQ water, 3 µl 20 µM universal adaptors and

10 µl Blunt/TA ligase master mix (NEB M0367S), or 50 µl 1× Quick ligation buffer, 3µl 20 µM adaptors and 2 µl NEB DNA Quick ligase (NEB, M2200). Ligation was performed for 30 mins at RT with rotation. Then the beads were again washed twice with TWB, then once with mQ water or 10 mM Tris-Hcl pH 8.0, and then resuspended in 50 µl of the same solution, and frozen at -20°C. The choice of mQ or Tris buffer for resuspension depended on the current batch of streptavidin beads: certain batches inhibit PCR (Rao et al., 2014), and in that case beads were resuspended in water for more efficient removal of DNA from the beads. In case of PCR-inhibiting beads, the DNA was eluted from the beads by heating them at 98°C for 20 mins, and moving the supernatant that contains DNA to a new tube. The test PCR was performed using the Q5 DNA polymerase (NEB M0491L) according to the manufacturer's recommendations with addition of 2×SYBR Green solution to monitor PCR in real time, and with the Illumina indexed primers. Annealing was set to 65°C, and extension for 40 seconds. The amplification curve was obtained for each sample to chose the appropriate cycle number to stay within logarithmic growth range. It varied between 12 and 14 cycles between different samples and experiments. Then the preparative PCR was performed in 4-6 reactions with the chosen number of cycles. Reactions were then pulled, concentrated on an Amicon filter and purified on AMPure beads (Beckman Coulter A63882, obtained from the MRC HGU Technical services) with 0.8:1 beads:DNA ratio to remove primers and primer dimers. The fragments were then size selected on 1.5% agarose gel to retain fragments of 200-700 bp. These final Hi-C libraries were pooled together equimolarly after validating successful size selection on a

Bioanalyzer, then sequenced with low depth on an Illumina NextSeq 550 (75 bp paired end mid output mode) at the Edinburgh Wellcome Trust Clinical Research Facility (WTCRF) to check the quality. High quality libraries were then sequenced deeply on HiSeq 4000 at BGI.

1.3 Hi-C data analysis

2.1.1 Read mapping and generation of Hi-C maps

I analysed Hi-C data from reads to genome-wide matrices using the *distiller* pipeline implemented in *nextflow* (https://github.com/mirnylab/distiller-nf) using the Eddie3 High-Performance Computing Cluster of the University of Edinburgh. Briefly, it uses *bwa mem* to map data to the reference genome (mm9 in our case), then parses and filters the alignments using *pairtools* (https://github.com/mirnylab/pairtools), and creates genome-wide matrices of interactions in the *Cooler* format (https://github.com/mirnylab/cooler) (Abdennur and Mirny, 2020). I used the pipeline with default settings (except max_mismatch_bp: 0, instead of 3) and, unless specified otherwise, used output files filtered for reads with mapping quality (mapq) >30 with iterative correction (balancing).

2.1.2 Quality control of Hi-C data

I used a custom script to analyse the output of distiller which creates plots that help with assessing library quality. It is available online on the GitHub Gists platform: https://gist.github.com/Phlya/9af1ffde527afe51e0558eb35e0025c7. In particular, I analysed the fraction of intra-chromosomal (cis) reads, the dependency of contact probability on distance, and PCR/optical duplicate fraction. I checked that the inter-chromosomal (trans) read fraction didn't exceed 20%, that the contact decay with distance was consistent between replicate samples, and the duplicate fraction was low to support deeper sequencing of the libraries.

2.1.3 Compartment and insulation analysis

I used *cooltools* (https://github.com/mirnylab/cooltools) to generate genome-wide tracks of insulation and of the first eigenvector. For insulation analysis, I used *diamond-insulation* (25 kb resolution data with 1 Mb window size for WT/ RING1B I53A/RING1B KO analysis, and 10 kb/100kb for the serum/2i analysis, since changes there are more subtle), and analysed the log_2 insulation score. I removed all bins which were filtered out during the balancing as coverage outliers. Similarly, I used *call-compartments* with 200 kb resolution data to find the whole-genome eigenvector that reflects the compartment structure. As a reference track to choose the best eigenvector among the top three and to orient its sign to get positive correlation with active chromatin, I used the track of GC content. For both analyses, for clustering I used the *seaborn* python package (Michael Waskom et al., 2018) with Euclidean distance and single linkage. For principal component analysis (PCA) I used the *scikit-learn* python package.

2.1.4 Annotation of chromatin loops

I used the deep Hi-C data from mESCs (Bonev et al., 2017) to find regions of locally enriched interactions, corresponding to chromatin loops, since our Hi-C data were not sufficiently deeply sequenced to identify the expected number of loops. We used *cooltools call-dots* reimplementation of the HiCCUPS algorithm (Rao et al., 2014) from branch *dekkerlab/shrink-donut-dotfinder* (commit 377106e). This was used with default settings (except for lower FDR threshold of 0.1, and 20 Mb as the maximal allowed distance separation between potential loops) with mESC Hi-C data at 5 kb, 10 kb and 25 kb resolution. Calls

from different resolutions were combined using a custom script following the HiCCUPS merging procedure (https://gist.github.com/Phlya/340af6c45900902310a7bb8d9cc60537).

Annotated dots were then filtered by intersecting with published CTCF peaks (Bonev et al., 2017), and/or RING1B peaks (Illingworth et al., 2015) using *bedtools pairtobed* after widening the peaks using *bedtools slop*. These loop annotations were then used to quantify loops strength in our lower coverage datasets.

2.1.5 Statistical testing of differential interaction frequencies

For a set of regions of interest where we visually observed changes in interaction frequency in Hi-C data, I performed statistical testing to make sure these changes are not observed by chance. To do that, I obtained observed/expected ratios for the regions of interest for both conditions I was comparing, and calculated average level of interaction enrichment. Then I obtained the same average values for 10,000 random regions of the same shape and size from the same chromosome (the same distance away from the main diagonal of the matrix). I then used these values to estimate the mean and the standard deviation of the distribution of all regions of the chromosome. Since after log transformation these distributions looked very similar to normal, I got the Z score for the region of interest by subtracting the mean from the observed value for the ROI and dividing by the standard deviation, and then converted it into a p-value for ease of interpretation (as

1-scipy.special.ndtr(zscore)). The code is presented here: https://gist.github.com/Phlya/27ccac1046e28e874d15c8273d540d33

2.1.6 ChIP-seq analysis and peak classification

Analysis of ChIP-seq data was performed by Rob Illingworth. Using published ChIP-seq data (Illingworth et al., 2015), peaks separated by less than 5000 bp were merged using *bedtools mergeBed*. Coverage of ChIP-seq reads (RING1B (Illingworth et al., 2015), CBX2 (Deaton et al., 2016), Mel18 (also known as PCGF2) (Morey et al., 2015), RYBP (Rose et al., 2016), KDM2B (Blackledge et al., 2014)), H3K27me3 (Illingworth et al., 2015) over these peak regions was then calculated using HOMER. For pileup analysis, peak regions were then split into quartiles by occupancy of RING1B, ratio of CBX2/RYBP signal, Mel18/KDM2B signal, or length. CGIs (Illingworth et al., 2010) were classified as RING1B or H3K27me3 positive, if they overlapped with RING1B or H3K27me3 peaks (Illingworth et al., 2015) (using *bedtools intersect -wa -u –a*).

Similarly, using these peak regions, I quantified levels of RING1B binding in serum and 2i culture conditions using published RING1B ChIP-seq data (Joshi et al., 2015) using *bedtools map*.

Similarly to the peak regions annotation, read count in 25 kb windows was found for RING1B ChIP-seq data for comparing with Hi-C contact frequency.

2.1.7 Local compaction analysis

To analyse local compaction, I calculated the amount of observed/expected contacts in all 25 kbp abutting windows along the genome, excluding the first two diagonals, and not counting windows with any missing data. I used

quantiles of RING1B ChIP-seq read counts in the same 25 kbp windows to split the Hi-C read counts into groups with varying levels of PRC1 binding, and compared the mean interaction frequency across groups and across conditions. The plots were made with *seaborn pointplot*, which also performs estimation of 95% confidence intervals using bootstrapping.

2.1.8 Genome-wide pileup analysis

Since this was one of the main methods I used during my PhD, I created a convenient tool to perform pileups which implements all options I needed – *coolpup.py* (a command-line interface tool written in **Py**thon to **p**ile-**up** Hi-C data stored in the **.cool** format (Abdennur and Mirny, 2020)). The tool is described in Chapter 3 in detail. It was largely used with default settings (except normalization to expected values instead of randomly shifted controls, and other options where stated) in the appropriate mode for each analysis.

2.1.9 Analysis of loop-ability

I calculated loop-ability using *coolpup.py* for all RING1B ChIP-seq peak regions (see the *Coolpup.py* chapter for details of this approach). As the outcome values, I used Enrichment3 values corresponding to the average enrichment of interactions in the central 3×3 square of the pileups. As predictors, I used read density from published ChIP-seq experiments for a variety of PRC1 components (see 2.1.6), and peak length. I performed linear modelling using these values after normalizing the predictors to a standard scale to ensure that the coefficient values are comparable. I used *scikit-learn* to fit a linear model and extract model coefficients, and generate p-values for all of

them (https://stackoverflow.com/a/27975633/1304161). Coefficient values were plotted as bars to compare them between predictors and between conditions.

Using loop-ability values in WT and KO cells, I selected RING1B peak regions that were on average interacting with other RING1B peak regions in RING1B-dependent manner (>1.5 Enrichment3 in WT and >1.5 fold change in WT over KO; values chosen arbitrarily, but they approximately correspond to reciprocal of 1% percentiles for both distributions). I analysed properties of these "loopers" in regard to level of binding of PRC1 components, H3K27me3 and length (see 2.1.6) using boxplots, and Mann-Whitney test to check significance of observed differences. I then compared loopers with "retainer" genes identified in previous studies of PRC1 function (Blackledge et al., 2019; Fursova et al., 2019). RefSeq IDs of these genes were kindly provided by Nadezda Fursova. I converted these IDs into mm9 coordinates using the RefSeq gene build and identified their transcriptions start sites (TSSs). I then found corresponding RING1B peak regions using *bedtools closest* and filtered them to limit distance from the TSS to under 1,000 bp upstream and 200 bp downstream. These peaks then were used to directly compare with the loopers I identified from Hi-C data.

2.2 ChIP-seq data analysis

For the analysis of CTCF and cohesin (SMC1α) binding in serum and 2i culture I used published data (Atlasi et al., 2019). I re-analysed it from sequencing data, because provided on GEO bigWig files were not normalized to sequencing depth. I used the `nf-core/atac-seq` pipeline (Ewels et al., 2019) for the analysis (I did not use a ChIP-seq pipeline because they require a control input sample, which was not available for this dataset). Briefly, the pipeline uses `trim_galore` for adapter trimming, bwa for mapping, `picard`, `samtools` and `bamtools` for read filtering, `bedtools` and `ucsc-tools` for generation of depth-normalised bigWig files and `MACS2` for peak annotation. I performed mapping to the mm9 genome assembly for compatibility with Hi-C data and other analyses, and used the default settings, except I set the `--narrow_peak` option for peak calling.

I then used `deeptools computeMatrix` to extract values surrounding the CTCF peaks (±1 kbp) for both motif orientations (see 3.3.1) from bigWig files, I then loaded this data in Python, generated average profiles and plotted the results.

Chapter 3: *Coolpup.py* – a versatile tool to perform pile-up analysis of Hi-C data

3.1 Abstract

Hi-C is widely used to investigate 3D genome organisation. However, a major limitation is the great sequencing depth required to detect looping interactions. An approach to mitigate this is genome-wide averaging (piling-up) of loops from high-resolution datasets, then measuring their prominence in less deeply sequenced data. We describe *coolpup.py* – a versatile tool for pile-up analysis of Hi-C data, demonstrate its utility by replicating findings regarding the role of CTCF/cohesin in genome organization, describe the dynamics of polycomb-mediated looping across cell cycle and investigate the effect of different data normalization strategies. A novel variation of the pile-up approach aids in statistical analysis of loops. *Coolpup.py* aids Hi-C analysis by allowing easy to use, versatile and efficient generation of pileups.

This Chapter is available as a preprint at https://www.biorxiv.org/content/10.1101/586537v3 and is planned for publication at a peer reviewed journal.

3.2 Rationale

Major advances in the study of 3D genome organization have come from the development of a family of Chromosome Conformation Capture (3C) methods (Dekker et al., 2002). While these all rely on the same principle of *in situ* proximity ligation of crosslinked and digested chromatin, the scope of each method varies depending on experimental processing and the method of quantification of the 3C library (Barutcu et al., 2016). Hi-C, a genome-wide 3C-derivative, is the method of choice to investigate the organization of the whole genome (Lieberman-Aiden et al., 2009; Rao et al., 2014).

One of the main challenges in Hi-C remains the required sequencing depth due to the extreme complexity of good quality Hi-C libraries. The output of Hi-C is a square matrix of interactions and therefore requires a vastly greater sequencing depth than most sequencing-based approaches that simply look for enrichment of reads linearly along the genome (Lajoie et al., 2015). This limits the resolution at which genomes can be analysed in 3D, since going beyond ~5 kbp resolution requires billions of read pairs for a mammalian genome.

Looping interactions are among the most interesting features that can be studied using Hi-C, including Hi-C. Chromatin loops bring distal regions in the genome into close proximity and are manifest in Hi-C data as foci of increased interaction frequency (Rao et al., 2014). The majority of loops identified in Hi-C data from mammalian cells correspond to CTCF/cohesin associated interactions, created by loop extrusion (Fudenberg et al., 2016; Gassler et al., 2017; Sanborn et al., 2015). CTCF/cohesin associated loops are closely related to topologically-associating domains (TADs), which in most cases are

encompassed in a loop, and which can in turn contain loops. TADs have been reported to constrain enhancer-promoter communication (Franke et al., 2016; Lupiáñez et al., 2015) and might be related to genome stability (Canela et al., 2017), while some loops have been suggested to correspond to enhancer-promoter contacts (Rao et al., 2014). In addition, distal polycomb sites can be brought together in 'loops' (Bonev et al., 2017; Joshi et al., 2015; McLaughlin et al., 2019).

To our knowledge, currently the only robust method to identify loops *de novo* requires very deep Hi-C libraries, on the order of over a billion Hi-C contacts (Rao et al., 2014). This means that the vast majority of Hi-C datasets cannot be used to identify loops. However, they can be used to quantify the average loop strength (i.e. enrichment of contacts in those loops relative to their local background). To do this one can average (or "pile up") all areas of the Hi-C maps containing loops, annotated in a high-depth dataset (Rao et al., 2014). This idea is very similar to "aggregate profiles" used, for example, in chromatin immunoprecipitation and sequencing (ChIP-seq) analysis to quantify signal in a subset of regions, except in Hi-C this is for a 2D matrix instead of a linear track. The same approach can of course be applied directly to the data where the loops were annotated. Apart from quantifying the strength of known features, the same analysis can be used to investigate whether certain regions, defined for example based on ChIP-seq peaks, tend to interact with each other on average. To our knowledge the first ever application of pile-up-like analysis was used to investigate clustering of pluripotency factor binding sites in mouse

embryonic stem (ES) cells (de Wit et al., 2013). Pile-up analysis can aid in the discovery of novel drivers of interactions.

Another challenge is that Hi-C is a population-based method, and only provides population average measurements. Several single-cell Hi-C approaches have been published (Flyamer et al., 2017; Nagano et al., 2013, 2017; Stevens et al., 2017; Tan et al., 2018; reviewed in Ulianov et al., 2017), however none of these provides data depth or resolution comparable to that which can be obtained from a population of thousands of cells (Díaz et al., 2018): the resulting matrices are too sparse to analyse individual regions, and only aggregate genome-wide metrics can be efficiently employed. Approaches to analyse strength of loops, TADs and genome compartmentalization from such data genome-wide have been developed (Flyamer et al., 2017). These are all based on the “pile-up” approach described above using data from single cells for the regions corresponding to specific features identified in population Hi-C, to boost the amount of reads used in the analysis.

Since its inception in the current form (Rao et al., 2014), originally termed APA (“Aggregate Peak Analysis”), pile-up analysis has been used in numerous publications, both to analyse single-cell Hi-C data (Flyamer et al., 2017; Gassler et al., 2017; Nagano et al., 2017) and as a general way of quantifying feature strength (Abdennur et al., 2018; Bonev et al., 2017; Díaz et al., 2018; Fudenberg et al., 2016; Hsieh et al., 2019; Krietenstein et al., 2019; Kruse et al., 2019; McLaughlin et al., 2019; Nora et al., 2017; Rao et al., 2017a; Rowley et al., 2019; Schwarzer et al., 2017). A visual interactive tool aimed to semi-manually classify and pile-up predefined regions has also been developed

(Lekschas et al., 2018). However no single computational tool can perform all the various kinds of pile-up analyses that have been used in the literature, including local and rescaled (features of different size or shape are averaged, e.g. average TADs) and off-diagonal (e.g. average loops) pile-ups with different normalization strategies (Table 1). At the same time, performing detailed analysis of Hi-C data remains difficult for non-specialists due to the absence of easy to use tools.

Here we present a unified command-line interface tool written in **Py**thon to **p**ile-**up** Hi-C data stored in the widely used and versatile **.cool** format (Abdennur and Mirny, 2020) (*coolpup.py*). A simple script for plotting the output of *coolpup.py* is provided in the package (*plotpup.py*), although for higher flexibility we suggest directly using *matplotlib* or another library.

Here we have applied *coolpup.py* to published data to investigate the effect of different normalization strategies on the resulting pileups, and to replicate published results to verify *coolpup.py*'s algorithm. We also present a novel variation of the pileup approach implemented in *coolpup.py* that retains some of the locus-specific information and would allow more detailed statistical analysis of looping interactions in Hi-C data. Using published single-cell Hi-C, we also investigate the dynamics of polycomb-associated looping revealing a different dynamics of looping across the cell cycle compared with CTCF loops.

Feature	Juicer	HiCExplorer	GENOVA	coolpup.py
Aggregate loops	+	-	+	+
Aggregate region pairs	-	+	+	+
Interactions between two region sets	-	+	-	+
Local pileups	-	+	-	+
(Local) rescaled pileups	-	-	+	+
Distance normalization	-	Expected (and z-score)	Fixed shifts (for pairwise analysis)	Expected or random shifts
Coverage normalization	-	-	-	+
Anchored pileups/loop-ability	-	-	-	+
Command Line Interface	+	+	-	+
Simple text output of pileups	+	+	-	+
Hi-C file format	.hic	.cool, .h5, other?	HiC-pro	.cool

Table 1: Comparison of tools to perform pileup analysis on Hi-C data.

Comparison of four tools for pileup analysis across a set of features: Juicer Aggregate Peak Analysis (APA) (Rao et al., 2014), HiCExplorer (hicAggregateContacts and hicAverageRegions) (Ramírez et al., 2018), GENOVA (APA, ATA and PE-SCAn) (Weide, 2019) and *coolpup.py*.

3.3 Methods

3.3.1 Sources of datasets & data analysis

As a proof of principal, we applied *coolpup.py* to publicly available Hi-C data (Bonev et al., 2017; Nora et al., 2017) using *distiller* (https://github.com/mirnylab/distiller-nf) to obtain *.cool* files filtered with a map quality (mapq) of ≥30. We used this data at 5 kbp resolution. In addition *.cool* files for single-nucleus Hi-C (snHi-C), together with coordinates of loops and TADs used in the original publication (kindly shared by Hugo Brandão) (Gassler et al., 2017; Rao et al., 2014), were reanalysed at 10 kbp resolution (without balancing and with coverage normalization and 10 random shifts). We also used single-cell Hi-C data for diploid serum-grown mouse ES cells from (Nagano et al., 2017) (*.cool* files were kindly shared by Aleksandra Galitsyna) at 5 kb resolution. We created pile-ups for each cell in the same manner as for snHi-C, and used average value of interactions in the 3×3 pixel square in the centre to get level of interaction enrichment. RING1B and H3K27me3 ChIP-seq peaks were taken from (Illingworth et al., 2015) and lifted over to the mm9 mouse genome assembly. The coordinates of biochemically defined CpG islands were taken from (Illingworth et al., 2010). CTCF ChIP-seq peaks were taken from (Bonev et al., 2017) and, following liftOver to the *mm9* assembly, intersected with CTCF motifs found in the mm9 genome using Biopython's *motifs* module (Cock et al., 2009). A human CTCF position-frequency matrix was downloaded from JASPAR (MA0139.1). We used only motifs with a score >7, and discounted peaks containing >1 motif.

Regions of high insulation (meaning low number of contacts crossing this regions) in the Bonev et al. data were called using *cooltools diamond-insulation* from 25 kbp resolution data and a window size of 1 Mb. The output was filtered to exclude boundaries with strength <0.1, and then pairs of consecutive boundaries were combined to create an annotation of TADs. TADs longer than 1,500 kbp were not used due to their likely artefactual nature (based on both visual inspection, and the fact that TAD sizes are reported to be on the order of a few hundred kbp in mammalian cells (Rao et al., 2014)). The same loop annotations for mouse embryonic stem (ES) cells were used as in our recent preprint (McLaughlin et al., 2019).

All figure panels were created using *matplotlib* (Hunter, 2007) and assembled in Inkscape.

3.3.2 *Coolpup.py* implementation

Coolpup.py is a versatile tool that uses *.cool* files as the main input together with a bed (chrom, start, end) or pairbed (chrom1, start1, end1, chrom2, start2, end2) file to define the regions under investigation. The tool is implemented as a *python* package which parses all arguments via *argparse*, performs the computation and saves the output file(s). It leverages the scientific python environment, taking advantage of *numpy* (Walt et al., 2011), *scipy* (Jones et al., 2001) and *pandas* (McKinney, 2010). A separate CLI tool included in the package (*plotpup.py*) can be used to visualize the results, and uses *matplotlib* (Hunter, 2007). The code is available on github (https://github.com/Phlya/coolpuppy) and the package can be installed using

pip, which then makes *coolpup.py* and *plotpup.py* available in the command line. Alternatively, all main functions can be accessed directly from *python*.

The overall procedure for piling up a lot of small regions is the following. To minimize the number of file reads (at the cost of required computer memory), a sparse representation of each chromosome's Hi-C contact matrix is loaded into memory. Then, using an iterator, each required location (on- or off-diagonal) is individually retrieved to generate a corresponding submatrix from the data (with some specified padding around the centre of the region of interest), and added to the matrix of the same shape, initialized with zeros, while keeping track of the number of summed up regions. If specified, coverage of the window on each side is recorded. Similarly, if needed, the window (and the coverage) is rescaled to a required shape. This is done for all chromosomes (optionally, in parallel using *multiprocessing*), and then all of the results are summed and then divided by the total number of windows. If specified, coverage normalization is applied at this stage. Then, unless otherwise specified, a normalization to remove distance-dependency of contact probability is applied. In most cases the best and most efficient way is to use a (chromosome-wide) expected value for each diagonal of the matrix, which can be obtained for a cooler file using, for example, *cooltools compute-expected*. With the assumption that the probability of interactions only depends on distance, the whole-chromosome expected matrix is diagonal-constant matrix A with diagonal values d (also known as a Toeplitz matrix), such as: $A_{i,j} = A_{i+1,j+1} = d_{|i-j|}$ The simplicity of this expected model allows trivial creation of a matrix containing expected values for an arbitrary region of the intra-chromosomal Hi-C map without generating the

whole matrix to avoid high memory requirements, which is done for each region of interest. All expected matrices are averaged to generate a normalizing matrix. Alternatively, if the expected values are not available, for example for single-cell Hi-C data, this normalization can be performed using randomly shifted control regions. In that case, to generate the normalizing matrix, the whole pile-up procedure is repeated, but the coordinates are randomly shifted. In the end, the resulting matrix of averaged ROIs is divided element-wise by the normalizing matrix to remove effects of distance.

If not specified, balanced data with chromosome-wide expected normalization was used when creating pileups, except for the zygote and single-cell Hi-C datasets, where randomly shifted controls and coverage normalization were used instead. For the single-cell Hi-C (Nagano et al., 2017) analysis we only used pairs of RING1B and convergent CTCF peaks within 100-800 kbp of each other, since previous analysis (data not shown) indicated this as the distance range where both looping modes are observed. For plotting average TADs, apart from observed/expected pileups in Figure 3.2B, we show the matrices after re-introduction of slow decay of interaction probability with distance. We simply perform element-wise multiplication of the observed/expected matrix by a matrix of the same shape where each diagonal i has the value $i^{0.25}$, starting from i=0.

3.3.3 Performance profiling

Coolpup.py performance was tested on the University of Edinburgh Open Grid Scheduler cluster (Eddie3). We used Hi-C data for mouse ES cells from (ref. Bonev et al., 2017) and from (Nora et al., 2017) for testing. To generate the

required large number of coordinates for testing, we used coordinates of the B3 repeat from RepeatMasker track available from UCSC Genome Browser. For coordinate pairs, we used all pairs of convergent CTCF sites, described above. A separate job was submitted for each measurement, and the runtime of the *coolpup.py* call was recorded. Subsets of different sizes were generated using *coolpup.py*'s `--subset` argument. Where not specified, 4 compute cores were utilized. All measurements were performed 5 times. Plotted everywhere are actual measured runtime values, the line shows mean values and shaded area - ±95% confidence interval, using the *seaborn* plotting package (Michael Waskom et al., 2018).

3.4 Results

3.4.1 Different normalization strategies implemented in *coolpup.py*

Coolpup.py is a tool designed to create pileups from Hi-C data. Hi-C data can be normalized in different ways to remove either technical biases, or uninteresting (in this context) biological signal of the decay of contact probability with genomic distance. *Coolpup.py* provides ways to deal with both of these problems.

Hi-C data are usually normalized to remove systematic biases, such as GC-content or restriction site frequency (Yaffe and Tanay, 2011). *Cooler* implements a matrix balancing (visibility equalization) approach to remove all potential biases (Imakaev et al., 2012) and, when available, it is recommended to use balanced data for pile-ups. However, sometimes, for example in single-cell Hi-C, removing biases is impossible due to sparsity of data. Therefore, using unbalanced data is also an option in *coolpup.py*. However, because of the averaging of multiple regions during the pile-up procedure, the effect of biases can however be partially mitigated by normalizing the matrix by the coverage (i.e. the total number of contacts of the bins in the chromosome) of the averaged regions (Flyamer et al., 2017). As illustrated in Figure 3.1A for both CTCF and polycomb (RING1B) loops, this approach reduces coverage variability between bins and removes sharp crosses from the central bin that is present with unbalanced data. This normalization seems to slightly over-correct, i.e. the value of the central pixel is consistently somewhat lower than

when using balanced data. However, the results overall look more similar to balanced data than without coverage normalization.

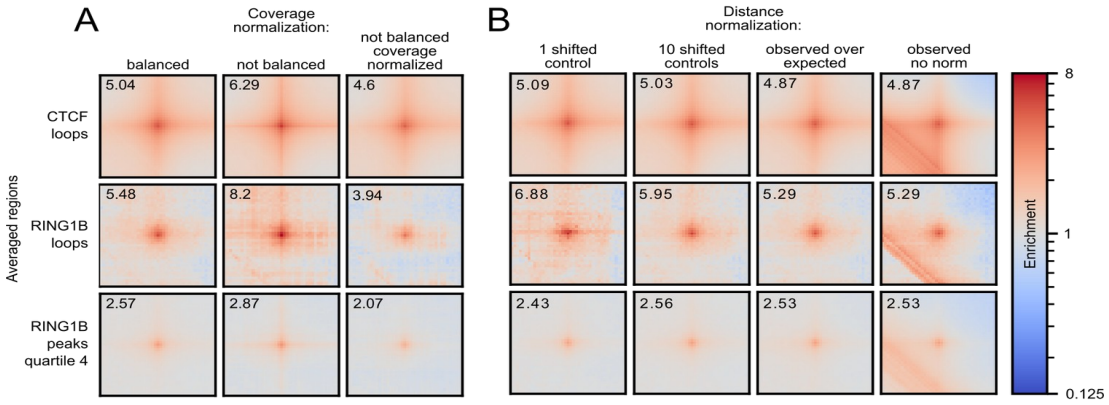

Figure 3.1. Hi-C data normalization strategies. (A) Comparison of coverage normalization strategies for pile-up analyses using mouse ES cell Hi-C data (Bonev et al., 2017). Normalization approaches are in columns: matrix balancing (iterative correction); no normalization; no balancing with coverage normalization of the pileups. The different averaged regions are shown in rows: loops associated with CTCF (n=6536), loops associated with RING1B (n=104) (see Methods), all pairwise combinations of high RING1B peak regions from the 4th quartile (by RING1B ChIP-seq read count) (n=2660 of peak regions). All pileups produced with 10 randomly shifted controls to remove short distance artefacts. All pileups are normalized to the average of the top-left and bottom-right corner pixels to bring them to same scale. Number in top left corner is value of the central pixel. 5 kb resolution with 100 kb padding around the central pixel (therefore, the each side of the pileup corresponds to 205 kbp). Colour is shown in log-scale and shows enrichment of interactions. **(B)** Same as A, but for different approaches to remove distance-dependency of contact probability with balanced data. In columns: single randomly shifted control regions per ROI; ten randomly shifted control per ROI; normalization to chromosome-wide expected; no normalization. Same rows as in (A).

In addition to normalization to remove biases, it is often desirable to remove the distance-dependency of contact probability in Hi-C data, since it can have a very strong effect on the resulting pileup by introducing artefacts from very

short distances, and sometimes can obscure interesting properties, such as enrichment in the centre of the pileup. However it is worth bearing in mind that this normalization can also hide real signal in the data, such as enrichment of interactions in the lower left corner, observed for CTCF-anchored loops (data not shown). The general approach to perform this normalization is to create a vector of expected contact frequency, which usually corresponds to the averaged value of the Hi-C map at each diagonal per chromosome. A file with such information (for balanced data) can be obtained using *cooltools compute-expected and* then used in *coolpup.py* to normalize the pileups. However sometimes the expected information is unavailable, for example, in single-cell Hi-C it can be too noisy. In that case, an alternative approach to remove distance-dependency of contact frequency can be used: for each position in the Hi-C map being averaged, a matched set of randomly shifted control regions with the same distance separation is used (Flyamer et al., 2017). In this way, by creating many such control regions for each region of interest (ROI), it is possible to estimate the expected frequency of interactions even for sparse single-cell Hi-C data. As shown in Figure 3.1B, both of these approaches are excellent at removing artefacts resulting from short-range interactions in the pileups and produce visually indistinguishable results. However, for a small set of regions (e.g. RING1B associated loops) a higher number of randomly shifted controls for each ROI is required to prevent noise. We note that for local pile-ups (especially with rescaling; see below) random controls perform better than simple normalization to expected values (data not shown).

3.4.2 Applications of pile-ups

As well as the basic pile-up procedure, there are multiple variations built in to *coolpup.py* which are tailored to answer different biological questions. The following ones are trivial, but worth mentioning. For example, often it is desirable to restrict the minimal and/or maximal separation of analysed sites, either to remove short-range artefacts, or to analyse the distribution of enrichment signal across different distance scales. Only certain chromosomes might need to be included, or, with too many regions of interest, a random subset can be taken to speed up the computation.

A popular variation of the pile-up approach is “local” pile-up: an analysis which focuses on near-diagonal features. For example, we averaged regions of high insulation annotated in the deep ES cell Hi-C dataset to visualize insulation strength after Auxin-induced degradation of CTCF (Nora et al., 2017)(Figure 3.2A). In this case the pileups are performed in the same way as previous off-diagonal pileups, however the regions that are averaged lie on the main diagonal of the Hi-C map. A variation of this approach is local pileups with rescaling to analyse features of different size, for example, TADs (Flyamer et al., 2017). As an example, TADs, based on aforementioned regions of high insulation annotated in data from (Bonev et al., 2017), were averaged to visualize changes in local interaction strength upon CTCF degradation (Nora et al., 2017) (Figure 3.2B). Here all windows centred on regions of interest are rescaled to the same size, and then averaged.

Pileups are a particularly important approach to analysing very low depth datasets to uncover genome-wide average patterns, which are indiscernible

when looking at individual regions in such shallow data. Here we apply *coolpup.py* to reproduce results from a dataset comprising pooled data from a few single cells, to show a loss of loops and TADs in mouse zygotes lacking SCC1 (RAD21), the kleisin subunit of cohesin (Gassler et al., 2017). Since the material is so limiting and data are based on single cells, the total number of contacts in this dataset is very low: 4.8 and 9.2 million contacts in *Scc1*$^{+/+}$ and *Scc1*$^{-/-}$, respectively. However, we successfully performed pileups, both with "traditional" averaging of loops, and local pileups of TADs with rescaling, and observe the loss of both loops and TADs upon deletion of cohesin, comparable to the original study (Figure 3.2C).

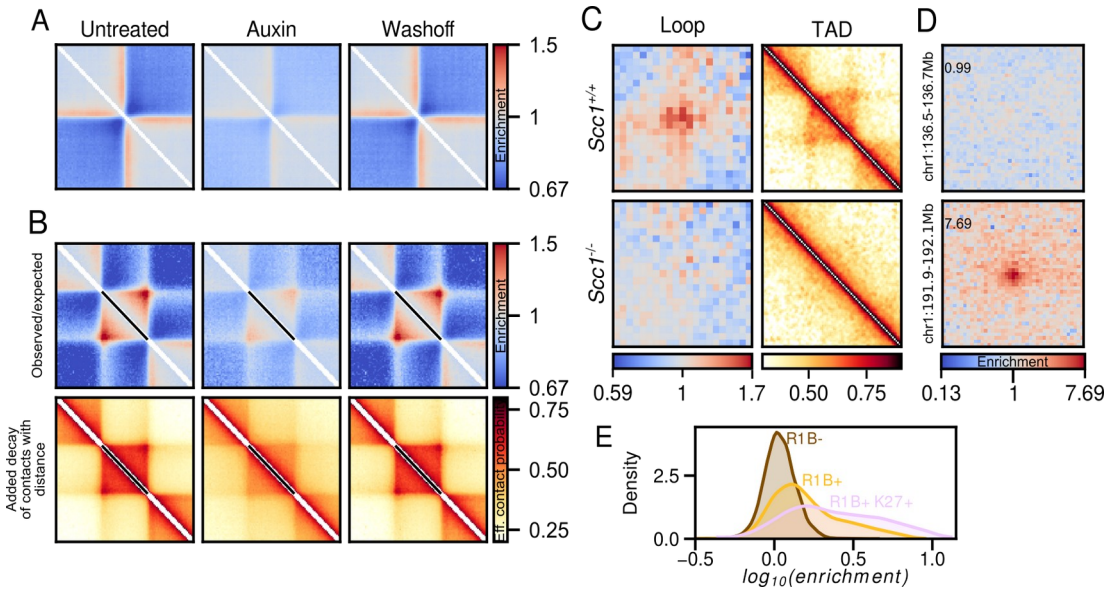

Figure 3.2. Pileup variations. (A) Local pileups of high insulating regions in ES cells across untreated, auxin-treated and wash-off conditions in CTCF-AID Hi-C data (Nora et al., 2017). 25 kb resolution data with 1,000 kb padding around the central pixel. **(B)** Local rescaled pileups of TADs (defined based on high insulating regions) across same data as in (A) from 5 kbp resolution data. Top row: raw observed over expected pileups; bottom row: pileups after artificial re-introduction of shallow decay of contact probability ($P(c)=s^{-0.25}$, see Methods) for easier visual interpretation of the data (Flyamer et al., 2017). **(C)** Loop and rescaled TAD pileups for pooled single-cell Hi-C data showing loss of structures in $Scc1^{-/-}$ zygotes (Gassler et al., 2017). **(D)** Two examples of anchored pileups from RING1B+/H3K27me3+ CpG islands, with no visible enrichment (top), or with very prominent enrichment (bottom). The anchored region is on the left side of the pileup, and its coordinates (including the padding) are shown on the left. The value of the central pixel ("loopability") shown in top left corner. **(E)** Distribution of "loopability" values of CpG islands not bound by RING1B, CpG islands bound by RING1B, and CpG islands bound by RING1B and also marked by H3K27me3.

All pile-up approaches include averaging of multiple regions, a drawback of which is loss of locus-specific information. We therefore designed a novel approach that retains some information about the specific loci used in the

analysis. In this approach, we pile-up a single region against multiple other regions; the same can be done for each of many regions in a set against all other regions. Then by extracting the value in the central pixel in pileups for each region, we can get a “loop-ability” value, which can then be related to other features of analysed regions, such as the level of occupancy by different factors. To confirm that this approach can work, we checked some example regions that displayed high or low level of “loop-ability”, to ensure that the values we observed were not due to noise from piling up interactions of a single region (see two examples in Figure 3.2D).

A simple proof of principle analysis highlights the interactions between sites bound by polycomb group proteins in mouse ES cells (data from Bonev et al., 2017). By splitting the CpG islands (data from Illingworth et al., 2010) (the main targets of polycomb binding in ES cells) in the mouse genome into RING1B (a core component of Polycomb Repressive Complex 1 - PRC1) negative, RING1B positive, and RING1B and H3K27me3 positive sets (data from Illingworth et al., 2015), we observe high “loop-ability” values for the two latter groups, while the RING1B negative CpG islands have close to no enrichment (Figure 3.2E).

Pileups are an invaluable tool when analysing Hi-C data from single cells, since averaging features across the whole genome helps to circumvent the sparsity of the data. Here we apply *coolpup.py* to analyse the looping interactions across the cell cycle using published single-cell Hi-C dataset from hundreds of mouse ES cells (Nagano et al., 2017). We compared the enrichment of interactions in different cell cycles stages for CTCF- and RING1B-associated

interactions (Figure 3.3A,B). For convergent CTCF sites, we detected the loss of loop strength in early G1, and in pre- and post-mitotic cells, consistent with the original publication (Nagano et al., 2017). In contrast, the interactions between RING1B binding sites have a very different dynamic across the cell cycle. They are at their weakest during S phase, progressively strengthening during G2 and not reaching their peak until early G1. This is consistent with the cell cycle kinetics of H3K27me3 abundance at polycomb marked sites with H3K27me3 levels lowest during S phase where they are diluted after the replication fork, with levels of H3K27me3 only accumulating slowly through G2 and not peaking again until G1 of the next cell cycle (Reverón-Gómez et al., 2018).

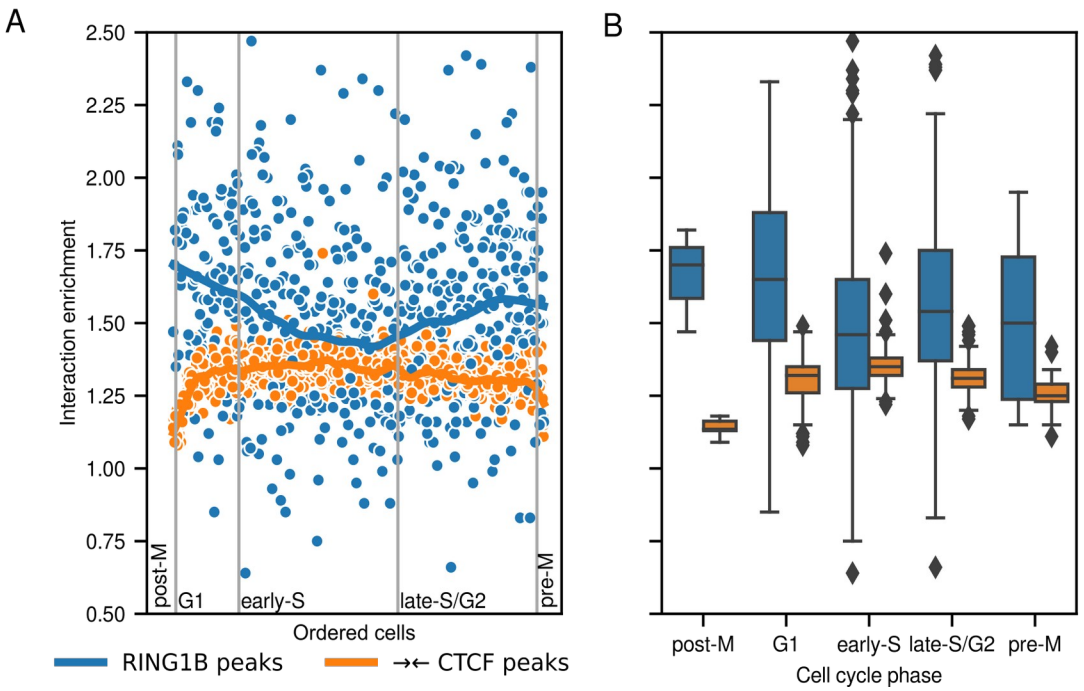

Figure 3.3. Chromatin looping dynamics across cell cycle. (A) Hi-C interaction enrichment levels for single cells ordered along the cell cycle (Nagano et al., 2017) for CTCF- and RING1B-associated loops. Curves represent LOWESS-smoothed data for easier interpretation. **(B)** Distribution of enrichment values in all cell cycle stages from data in (A).

3.4.3 *Coolpup.py* can deal with huge numbers of regions

Creating pileups from intersections of genomic regions can require averaging a huge number of 2D windows: the number of 2-combinations grows quickly with the number of regions. For example, with ~1000 regions per chromosome (which is approximately equivalent to the number of genes), requires averaging of ~10,000,000 windows for the whole genome, several orders of magnitude more than the number of regions usually averaged, such as number of annotated loops (~10,000). Therefore, it is important for a general-purpose tool for creating pileups to scale well with the number of averaged 2D windows. To facilitate this, *coolpup.py* performs a very low number of read operations on the

Hi-C data – only once per chromosome (or twice, when using randomly shifted controls). Whilst this necessitates that the whole Hi-C matrix of a chromosome has to be loaded into memory, it is only stored in a sparse format, and so conventional Hi-C datasets can be analysed on a regular desktop (although multi-billion contact datasets might require a high-memory machine; data not shown).

To test the performance of *coolpup.py* and how this depends on number of regions of interest, we measured the runtime with varying number of two-sided coordinate pairs (mimicking loop annotation) (Figure 3.4A), and varying the number of one-sided coordinate interactions being averaged (Figure 3.4B). We used both deep Hi-C data (Bonev et al., 2017), and "regular depth" data (Nora et al., 2017) from mouse ES cells. With both datasets, the runtime was almost constant up to a certain number of "loops" (~1-2×10^5), where it starts quickly increasing (Figure 3.4A). Notably, the best annotations that exist to date only contain <40,000 loops (Krietenstein et al., 2019), and therefore this would fall within the flat part of the curve. Similarly, in the latter analysis, runtime didn't increase up to 1600 regions of interest, and started growing quickly after reaching ~10,000 ROIs. Importantly, in both analyses the difference in time between datasets with almost 10-fold sequencing depth difference is not very large, and probably largely driven by differences in time required to read the data from disk.

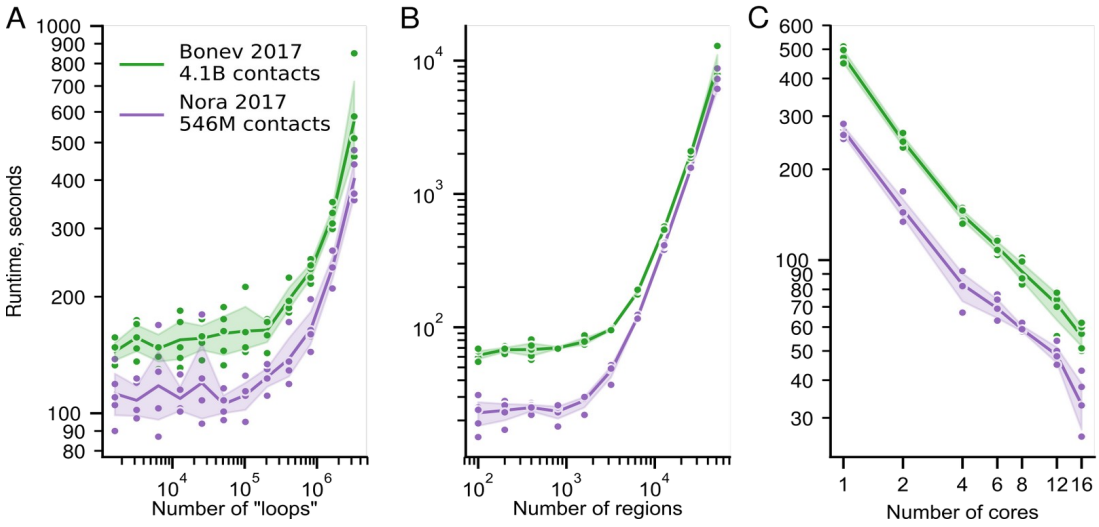

Figure 3.4. **Performance profiling. (A)** Runtime (seconds) of coolpup.py with varying number of averaged "loops" for two Hi-C datasets with different depth, using 4 cores. **(B)** Same as (A), but for number of linear regions between which interactions are averaged. **(C)** Runtime of the same analysis with 5000 linear regions and a varying number of cores.

Since *coolpup.py* supports parallel processing to speed up analyses, we also tested how well it scales with the number of computer cores used. We measured the runtime of the same analysis performed with varying number of cores (Figure 3.4C) and showed that the runtime shortened linearly with additional processes. This means the parallelization strategy used in *coolpup.py* is efficiently utilizing available CPU cores and when available, we recommend using many cores to speed up computation, although this would also significantly increase memory requirements.

3.5 Discussion

With the large efforts being made in deciphering the structure and function of the genome in 3D, efficient, robust and versatile tools are required to facilitate quick hypothesis testing. Unlike for RNA-seq, ChIP-seq and other genome-wide methods, analysis of complex Hi-C data remains a challenge only readily accessible to specialists in the field due to an absence of easy to use informatics tools, with a few exceptions. One popular analysis applied to Hi-C data is pile-ups, which show an average genome-wide view of a selected set of regions in the 2D Hi-C interaction matrix: a very visual and intuitive approach at analysing data.

Here we presented *coolpup.py*, a versatile tool to perform pile-up analysis on Hi-C data in *.cool* format. Apart from simple generation of pile-ups, *coolpup.py* can be used to explore different data normalization strategies. While we recommend using balanced data with normalization to chromosome-wide expected interaction frequency, in certain cases a different normalization strategy can be beneficial. Similarly, exploring other parameters of the algorithm (such as minimal separation between averaged loop bases, or minimal length of locally averaged features) is straightforward with *coolpup.py*. Using our tool, we reproduced published results on the role of CTCF and cohesin in generating chromatin loops and TADs. We have shown application of *coolpup.py* to both low coverage Hi-C data (merged snHi-C data), and extremely sparse single-cell Hi-C data. The latter analysis not only replicated published data on CTCF-mediated looping changes across the cell cycles, but also revealed novel cell cycle dynamics of polycomb-associated interactions of

highest contact enrichment around the time of mitosis. We note that these observations are generally consistent with the reduction and slow recovery of the H3K27me3 mark after replication (Alabert et al., 2015; Reverón-Gómez et al., 2018), as well as an antagonistic relationship between cohesin-mediated loop extrusion and looping between RING1B target sites, reported previously (Rhodes et al., 2020)(Rhodes et al., 2020)(Rhodes et al., 2020). These observations also pose a question whether polycomb-associated interactions persist in metaphase chromosomes - a possibility since components of CBX2-containing PRC1 remain associated with metaphase chromosomes (Zhen et al., 2014). These novel insights highlight the exploratory power of such pile-up analysis.

Since *coolpup.py* is designed as a command-line tool and allows reading the coordinates of regions from standard input, it is compatible with computational pipelines, and can be readily used in shared computing environments. Moreover, it remains accessible for non-specialists with minimal knowledge of the command line and no programming experience. *Coolpup.py* should aid in improving reproducibility by providing a standardised approach for pile-up analysis which is intuitive and therefore accessible to both specialists and non-specialist alike. We hope that it will facilitate research into the 3D organization of the genome by allowing easy to use, versatile and efficient generation of pileups.

3.6 Conclusions

Coolpup.py is a versatile tool able to perform a variety of pile-up analyses. It takes widely used file formats as input, can be applied to large datasets and included in computational pipelines. It is easy to use by non-specialists, and should make Hi-C data analysis more accessible to the field as a whole.

3.7 Availability of data and materials

Coolpup.py is available on GitHub (https://github.com/Phlya/coolpuppy, doi: 10.5281/zenodo.3237784) and can be installed from the Python Package Index (PyPI). It is distributed under the permissive MIT license. All code used to generate the results and figures for this paper is available on GitHub (https://github.com/Phlya/coolpuppy_paper).

3.8 Acknowledgements

We are grateful to Maksim Imakaev, Nezar Abdennur, Anton Goloborodko, Hugo Brandão and Sergey Venev for discussions, help and advice about the *cooler* ecosystem, and to Aleksandra Galitsyna for sharing *.cool* files for the Nagano et al., 2017 dataset. This work has made use of the resources provided by the Edinburgh Compute and Data Facility (ECDF) (http://www.ecdf.ed.ac.uk/). We also thank Sergey Ulianov for critically reading the manuscript.

Chapter 4: Canonical PRC1 folds the genome in 3D in mouse ES cells

4.1 Abstract

While currently it is not understood how PRCs silence genes, one of the main features of domains of PRC binding is altered chromatin structure. Specifically, targets of PRC1 are highly compacted, and cluster together in nuclear space (Eskeland et al., 2010; Joshi et al., 2015; Kundu et al., 2017; Schoenfelder et al., 2015). An appealing model is that this structural property of polycomb complexes contributes to gene silencing, perhaps by reducing accessibility of the promoters to activating factors, or their receptiveness to enhancer-driven activation (Kundu et al., 2017). It is also likely to increase the robustness of gene regulation by creating a larger repressive environment and high concentration of silencing factors.

In this Chapter I describe my findings about the properties of PRC1-mediated chromatin structures and the role of PRC1 subcomplexes in creating them in mouse ES cells, using Hi-C. Comparing wild-type cells to a RING1B knock-out line, I show that the PRC1 complex creates both local compaction of its targets, and distal interaction between them. I also used ES cells with a RING1B I53A point mutation which has greatly impaired E3 ubiquitin ligase activity (Buchwald et al., 2006; Elderkin et al., 2007; Illingworth et al., 2015). My data suggest that there is no direct role of the H2AK119ub mark deposited by RING1B in formation of 3D chromatin structures. I have gone on to computationally investigate what determines formation of distal interactions between RING1B binding sites. Surprisingly, distance turned out not to be an important factor, since I detect enriched interactions even at genomic separations of 50-100 Mbp. This is in contrast to loops mediated by CTCF which are restricted to a

distance scale up to ~1-2 Mbp. Moreover, consistent with locus-specific examples in the literature (Kundu et al., 2017), I could detect a clear role of canonical PRC1 binding, but not PRC2 or variant PRC1 complexes in forming these interactions. Further taking advantage of our genome-wide data,I show that transcriptional activation of polycomb targets in cells lacking RING1B is not required for loss of interactions between them.

The most important findings of this Chapter were included in our manuscript (Boyle, Flyamer, et al., manuscript in preparation). Some other findings by my collaborators included in the preprint will be referenced in the Chapter.

4.2 Results

4.2.1 General analysis of Hi-C data from RING1B mutant mESCs

To investigate the role of PRC1 in organisation of the genome in 3D genome-wide, I performed *in situ* Hi-C on three mES cell lines: wild type E14TG2a (denoted WT below), RING1B$^{-/-}$ E14 ES cells (denoted KO below), and RING1B$^{I53A/I53A}$ (denoted I53A below) (Illingworth et al., 2015). Since RING1B is the predominant RING1 homologue expressed in mESCs, knock-out of this gene essentially leads to loss of PRC1 function (Leeb and Wutz, 2007). RING1A can potentially partially compensate for loss of RING1B. However double-knockout cells lose ES cell-specific morphology and probably spontaneously differentiate, and therefore any differences are convoluted by overlay of loss of PRC1 and differentiation (Endoh et al., 2008). Levels of many PRC1 components also drop in RING1B knock-out cells, suggesting their degradation when not part of the multiprotein complex (Leeb and Wutz, 2007). I53A mutation disrupts the interaction between RING1B and the E2 ubiquitin ligases, while preserving interactions within the PRC1 complex (Buchwald et al., 2006; Elderkin et al., 2007; Illingworth et al., 2015). Therefore it causes a great loss (while not a complete loss of function) of the E3 ubiquitin ligase activity of RING1B, and H2A119Ub levels (Blackledge et al., 2019; Cohen et al., 2018; Illingworth et al., 2015). We used the I53A cell line to investigate the potential role of the H2AK119ub in directing genome folding. A severe depletion of ubiquitination levels in cells with this mutation would reveal its

involvement in 3D organization, while low level of residual modification still allows correct targeting of polycomb components (Blackledge et al., 2019).

I generated 4 independent Hi-C datasets for WT cells, and 2 datasets each for I53A and KO cells. I analysed the data using the *distiller*-nf pipeline to perform read mapping, filtering and binning into the Hi-C matrices. In total, I obtained approximately 300 million intra-chromosomal contacts longer than 1000 bp for each condition (Figure 4.1A). Quality of the data, as measured by the percentage of usable reads (~30-50% of total reads, except WT-3 replicate with 25% reads) was consistent between replicates and conditions.

One of the key measures used to assess the gross nuclear organization using Hi-C data, is analysis of $P_c(s)$ curves, which describe how contact probability changes with genomic separation (Lieberman-Aiden et al., 2009; Naumova et al., 2013). Changes in the slope of the curve can reveal changes in chromatin compaction at different scales. Knowing that the nuclear size is significantly increased in RING1B KO cells (or other conditions of loss of polycomb binding) while cell cycle profile is preserved (Boyle et al., 2019), I expected to observe a steeper $P_c(s)$ (Flyamer et al., 2017) in that condition. However, surprisingly I didn't observe any difference between conditions (Figure 4.1B) using this measurement. This suggests that the gross chromatin density is not significantly changed in either of the mutant cells, and expansion of nuclear size is not linked to general chromatin decompaction. One possible explanation for this is that the change in nuclear size is driven by the expansion of the interchromatin compartment (IC) with preserved volume occupied by chromatin, potentially due to presence of more RNA from derepressed genes in

KO cells. Since IC is present between the chromosome territories, I reasoned that if it's expanded, the frequency of inter-chromosomal (or *trans*) contacts should be diminished in KO cells. Therefore, I investigated the fraction of inter-chromosomal contacts in these datasets (Figure 4.1C). Surprisingly, the fraction of trans contacts was slightly higher in the data from KO cells. This is consistent with decompaction of chromosome territories, and does not agree with the $P_c(s)$ analysis. A potential explanation for this could be a specific expansion of IC in the interior of chromosome territories, which would not lead to chromatin decompaction, but would reduce the relative volume of the IC between chromosome territories. This could be caused by transcriptional derepression of polycomb targets upon RING1B KO. I also observed a very subtle loss of trans contacts in the I53A data, but its biological significance is unclear.

I next investigated more specific features of the Hi-C maps: chromatin compartmentalization (Imakaev et al., 2012) and local insulation (Crane et al., 2015). Compartmentalization is measured at low resolution (200 kbp here) by eigenvector decomposition of the Hi-C matrix and choosing one of the top eigenvectors with the highest correlation with GC content (Imakaev et al., 2012). This reveals the pattern of spatial segregation of active and inactive compartments in the nucleus, and therefore is affected by the transcriptional state of the cells. Since transcriptional changes in RING1B KO cells are much greater than in I53A cells (Illingworth et al., 2015), I expected to see a bigger difference for KO from the WT cells.

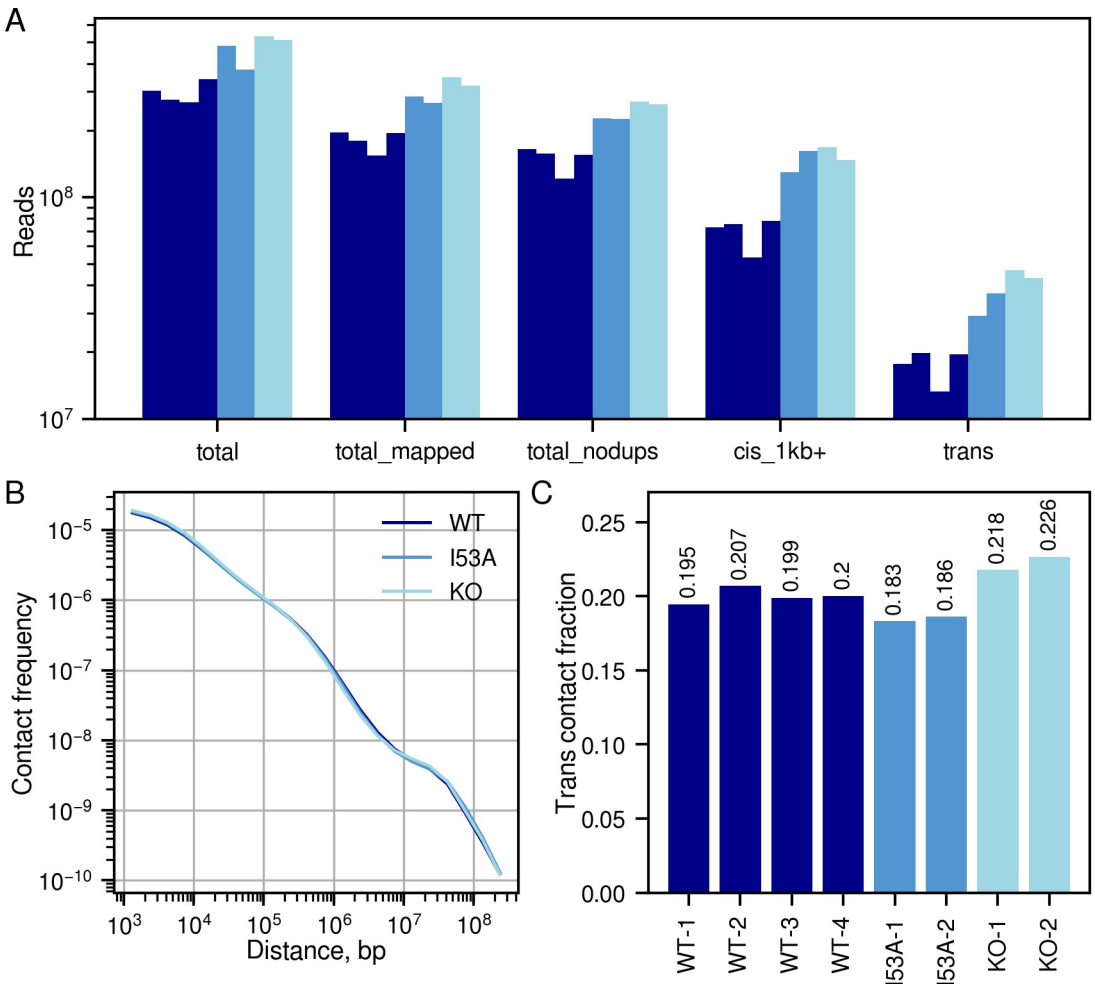

Figure 4.1. **Quality control and comparison of gross properties of Hi-C data.** **(A)** Read statistics of Hi-C datasets used in this Chapter: four replicates of wild type ES cells (dark blue), two replicates of RING1B$^{I53A/I53A}$ cells (blue), and two replicates of RING1B$^{-/-}$ cells (light blue). Y axis shows number or reads, and is log-scaled. Categories are: number of reads ("total"), number of mapped reads ("total_mapped"), number of mapped reads after removal of PCR/optical duplicates ("total_nodups"), number of intra-chromosomal contacts longer than 1 kbp ("cis_1kb+"), number of inter-chromosomal contacts ("trans"). **(B)** $P_c(s)$ curves for combined data for three conditions used in this Chapter. **(C)** Fraction of inter-chromosomal contacts in all replicates of the data used in this Chapter.

Consistent with the massive gene derepression in RING1B KO (hundreds of genes (Illingworth et al., 2015)), KO datasets clustered separately from the WT

and I53A cells (Figure 4.2A). At the same time I53A cells clustered together as a subgroup of the WT, consistently with the minor transcriptional changes in this condition (Illingworth et al., 2015). Looking at the same data using Principle Components Analysis (PCA), I observed a similar pattern (Figure 4.2B) with KO replicates visibly separated from the rest of the data. However, WT cells are very spread out along the second principle component (and other top 6 components, with the main pattern preserved; data not shown). This is probably due to the ~50% lower sequencing depth of all WT replicates relative to the I53A and KO replicates, which causes higher variability in compartment estimation.

Next I performed a genome-wide analysis of local insulation using insulation index (Crane et al., 2015). This simple approach counts how many contacts cross a particular genomic location within a restricted distance window, and can be performed at a high resolution. Here I used 25 kbp resolution and 1 Mbp window size (chosen by visual inspection of insulation peak calls, however results are similar with other parameters; data not shown).

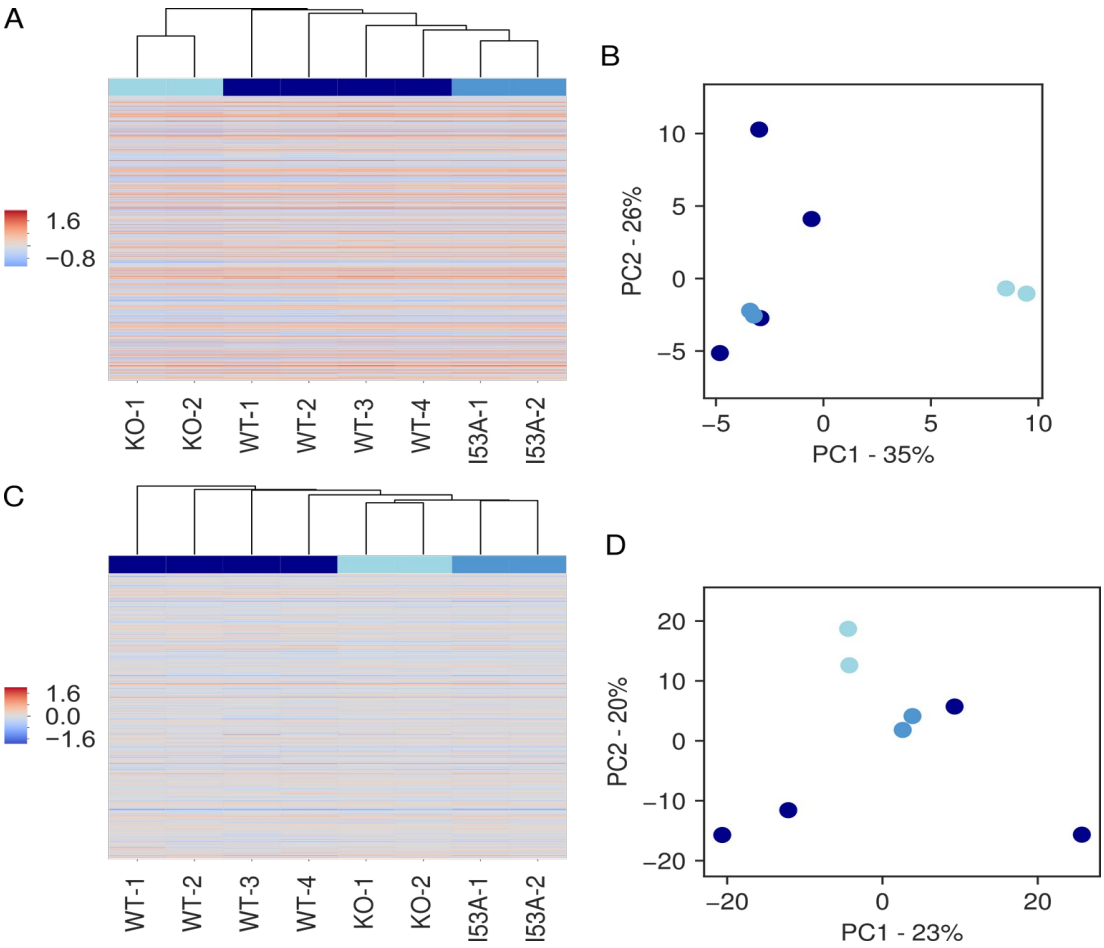

Figure 4.2. Comparison of compartmentalization and local insulation. (A) Hierarchical clustering of genome-wide compartment signal (first eigenvector) of Hi-C data from all Hi-C data used in this Chapter (200 kbp resolution). **(B)** Principle component analysis of data from (A). Colours of the points indicate the genotype. **(C)** Same as A, but for insulation index (25 kbp resolution, 1 Mbp window size). **(D)** Principle component analysis of data from (C). Colours of the points indicate the genotype.

I analysed the results the same way as above for the compartmentalization analysis, using hierarchical clustering (Figure 4.2C) and PCA (Figure 4.2D). In the clustering I observed KO and I53A cells form separate subgroups together within the larger cluster of WT cells. Looking at the PCA plot, it is clear that the

situation is very similar to the compartmentalization analysis with separation of KO cells from the rest, except the variability of WT replicates is higher and causes less clear separation of other conditions from them. However, the smaller effect of RING1B loss on local insulation compared to compartmentalization is consistent with much more subtle local chromatin structures formed by PRC1, than compartmentalization dependent on gene silencing mediated by polycomb. Of note, performing clustering or PCA analysis on combined data recovers the expected structure of KO data being a conspicuous outlier relative to the WT and I53A cells, which are clustered together for both discussed metrics (data not shown), but significance of such analyses with just 3 groups is questionable.

4.2.2 Compaction of long PRC1 targets

Certain polycomb binding targets have been shown to be highly compacted in ES cells, with this compaction dependant on PRC1 binding (Eskeland et al., 2010; Kundu et al., 2017; Williamson et al., 2014). Most of the research in this area has focused on the four paralogous Hox loci, since these Hox gene clusters are very long (~100 kbp) and have particularly high levels of PRC1 binding (Schoenfelder et al., 2015). However it has also been shown, by high-resolution 5C on a few examples, that other regions of high PRC1 binding form similar structures with increased level of interactions inside them (Kundu et al., 2017). Here I address this question genome-wide using my Hi-C data.

First, I visually confirmed that I observe similar patterns of increased interaction frequency at Hox loci. Shown in Figure 4.3A-D are all Hox clusters, and they all clearly form highly compacted structures with a lot of local interactions in Hi-C

data, visible as triangles of darker colour in the Hi-C map. In comparison, there is a very subtle decrease in interaction frequency in the PRC1-bound region in the data from I53A cells. However, the investigation of the same regions in data from RING1B KO shows complete loss of these dense structures. This is observed across all four Hox regions, and it is consistent with previous reports (Eskeland et al., 2010; Kundu et al., 2017). I tested these changes for statistical significance by comparing them to the distribution of observed/expected signal ratios of 10,000 random regions of identical size from the same chromosome (Figure 4.3EF, Table 2) . Interestingly, the decrease of interactions in I53A cells compared to WT cells is significant with a Z-score of 1.9 for the HoxD, but the rest of Hox loci don't change significantly in this condition. However, when comparing data from KO cells to WT cells, all Hox regions display a high level of significance with Z-scores of ~8. Exact Z-scores and p-values are available in Table 2. A potential reason for this difference in decompaction in I53A cells could be that the HoxD and HoxC (the most affected of the other three Hox loci) are not split into a major and minor parts, while both HoxA and HoxB have a gap with a few genes separated from the rest. It is most prominent in HoxB, which does not change at all in I53A cells.

Next I looked at other extended regions of PRC1 occupancy in the genome (RING1B ChIP-seq data from Illingworth et al., 2015). I could observe local enrichment of interaction in all regions investigated earlier by 5C (Kundu et al., 2017). One example of such region is the *Cbx2/Cbx8/Cbx4* gene cluster, where the former two genes and part of the intergenic sequence are coated with H3K27me3 and RING1B (Figure 4.4A), forming a >50 kbp long region of PRC1

binding with a short gap between the two genes. A very similar pattern to the HoxD region is observed here, with a region of high interaction frequency in WT cells, subtle weakening in I53A, and a much greater loss in RING1B KO data. This is mirrored by the statistical analysis of this region, with Z-score of 1.4 in comparison of I53A cells with WT, which corresponds to a p-value of ~0.08 (Figure 4.4C, Table 2). In contrast, the change is highly significant in KO cells with Z-score of 4.07. Interestingly, the local enrichment of contacts is incompletely in this case, unlike for Hox clusters. Either it is maintained here by a different mechanism operating independent of PRC1 (e.g. cohesin/CTCF mediated loop extrusion), or in this region RING1A can partially compensate for loss of RING1B. Very similar results were obtained for another extended PRC1 target, the region containing the *Nr2f2* gene (Figure 4.4B, D, Table 2).

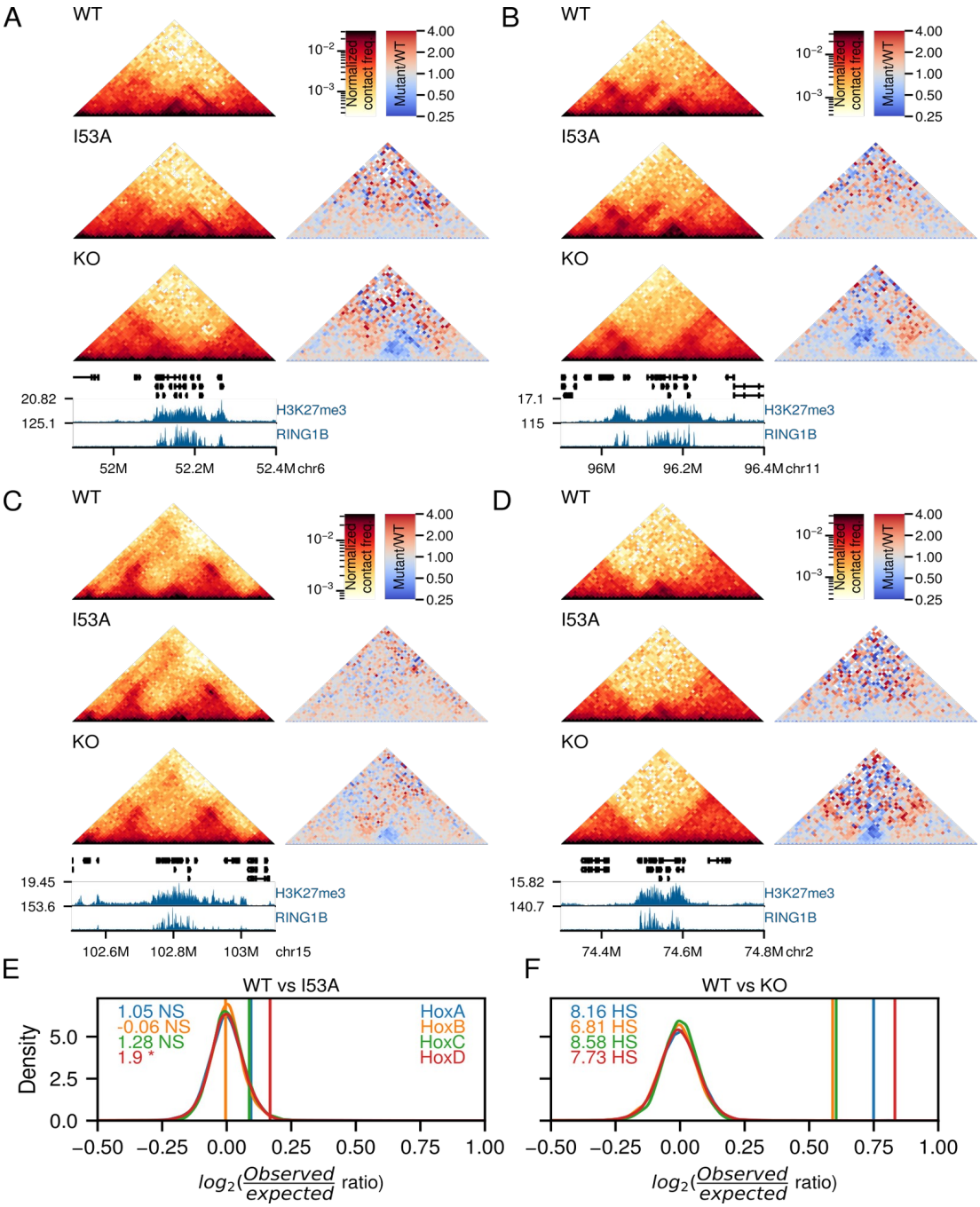

Figure 4.3. Local compaction of Hox clusters targets. (A, B, C, D) Hi-C maps at 10 kbp resolution of the HoxA/B/C/D gene clusters, respectively, in WT, I53A and KO cells, and differential maps of ratio of mutant data over WT data. Genes, H3K27me3 and RING1B ChIP-seq signals (from WT cells) are shown below. **(E, F)** Testing decompaction in the Hox regions. The curves show the distribution of observed/expected ratios for 10,000 random regions, lines – the ratios for Hox regions.

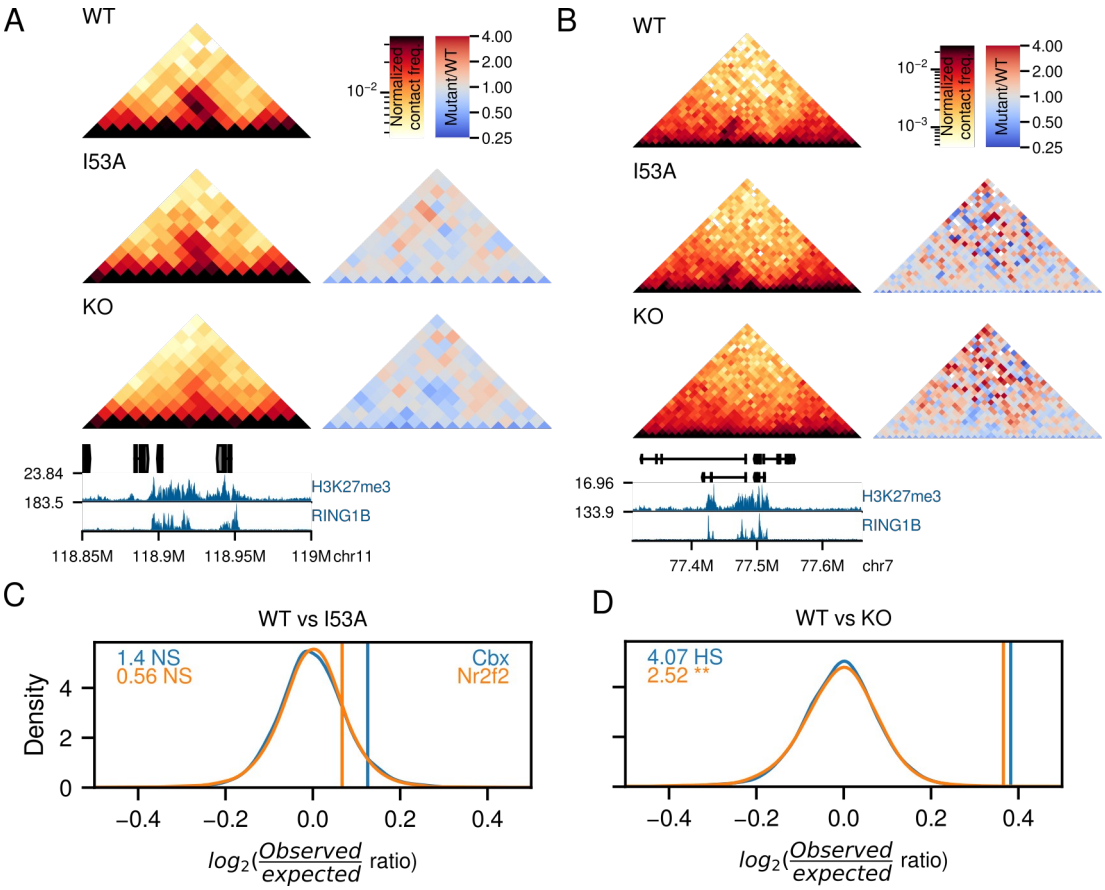

Figure 4.4. Local compaction of extended non-Hox Polycomb targets. (A, B) Same as Figure 4.3A-D, but for the *Cbx2/4/8* (A) and *Nr2f2* (B) gene regions. **(C, D)** Same as Figure 4.3E-F, but for the *Cbx2/4/8* and *Nr2f2* gene regions.

I then decided to quantify this loss of local interactions genome-wide. I used two approaches to do this. First, I used *coolpup.py* to create local rescaled pile-ups (see Chapter 2) of long continuous regions of RING1B binding. I chose all ≥10 kbp long regions with RING1B ChIP-seq peaks within 5 kbp of each other. There were 181 regions like that in the mouse ES cell genome. I used 1 kbp resolution Hi-C data, since the regions of interest were too short to analyse at lower resolution. Averaging parts of the WT Hi-C map corresponding to these regions, revealed a pattern of locally increased interactions which corresponded to the the RING1B binding (Figure 4.5A). As expected from the

above analysis of some example regions, in I53A cells this pattern is weakened, but it essentially completely disappears in RING1B KO cells. This shows that chromatin compaction of PRC1 targets is observed genome-wide for at least the extended regions of RING1B binding, and it depends on the presence of RING1B/PRC1. Whilst with this analysis it is impossible to tell what fraction of regions displays this pattern but it is sufficient to generate a clear enrichment after averaging all of them.

Name	Coordinates	I53A z-score	I53A p-value	KO z-score	KO p-value
Cbx	chr11:118880000-118955000	1.40	0.08	4.03	2.8E-05
Nr2f2	chr7:77425000-77520000	0.56	0.29	2.52	5.8E-04
HoxA	chr6:52100000-52220000	1.05	0.15	8.16	1.1E-16
HoxB	chr11:96110000-96215000	-0.06	0.52	7.61	1.4E-14
HoxC	chr15:102740000-102870000	1.27	0.1	8.66	0
HoxD	chr2:74495000-74605000	1.90	0.03	8.01	5.6E-16

Table 2: statistical analysis of changes in local compaction between WT and I53A or KO cells for regions discussed in this Chapter: four Hox regions, the Cbx gene cluster and the Nr2f2 region.

An alternative approach I used to quantify the same phenomenon was splitting the whole genome into 25 kbp windows of different levels of RING1B occupancy from ChIP-seq data (Figure 4.5B). Since the vast majority of the genome has very low level of RING1B occupancy as detected by ChIP, I used very unevenly sized groups. The first group of windows contained the 99.5% of the genome with essentially no RING1B binding. The rest was split into three

groups of increasing RING1B binding by the 99.8 and 99.9 percentiles. Then for each of these groups, I calculated the average number of observed/expected interactions within these windows, and plotted this together with 95% confidence interval error bars to quantify local chromatin interactions/ compaction across these groups. This approach revealed a lower than average (in the first group) level of compaction in the 99.5-99.8 percentile group. This is probably related to generally high gene density, since PRC1 and PRC2 are found at CpG islands (Ku et al., 2008), and compartment A association of RING1B bound regions in ES cells (data not shown), with insufficient level of PRC1 binding to detect compaction generated by it. In WT and I53A conditions, the level of interactions increases with higher RING1B binding, with slightly lower signal in the I53A data. In KO data, however, the level of interactions decreases with higher level of RING1B binding. This creates a very large difference in average interaction frequencies between WT and I53A data, and KO data for the highest 0.1% genome by RING1B binding, which corresponds to 97 windows (2.425 Mbp total). As in previous analyses, I53A cells show slightly lower level of interactions that WT here, but the compaction is still clearly observed relative to RING1B KO data.

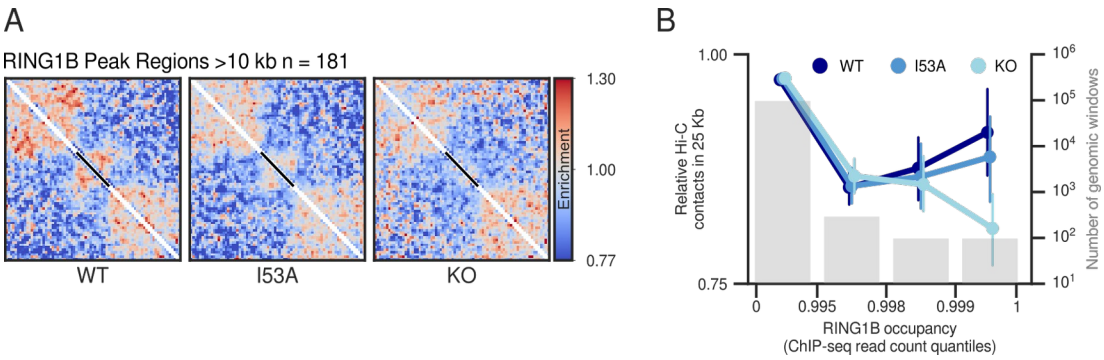

Figure 4.5. Genome-wide quantification of local compaction by PRC1. (A) Local rescaled pileups of all long regions of RING1B binding (length>10 kbp, n=181) in WT, I53A and KO Hi-C data. Black bar shows the location of the averaged RING1B binding sites. **(B)** Mean number of normalized local Hi-C interactions in 25 kbp windows of varying levels of RING1B binding from ChIP-seq (split into four groups), for WT, I53A and KO data. Error bars show 95% confidence interval obtained by bootstrapping. Number of regions in each category is shown as grey bars with value on the right Y axis.

4.2.3 Interactions between distal PRC1 binding sites

Apart from local compaction, distal interactions between targets of PRC1 binding have been previously reported (Bonev et al., 2017; Joshi et al., 2015; Kundu et al., 2017; Rhodes et al., 2020; Schoenfelder et al., 2015). I could clearly see these in my Hi-C data in a number of regions. Therefore, I decided to take advantage of my genome-wide data to study these in detail: previous analyses were limited by selection of regions for FISH/capture-C/5C, or didn't investigate why some regions interact with each other, while others don't.

One of the most prominent interactions I have seen is formed between two large PRC1 domains, which contain *Skida1* and *Bmi1* genes (Figure 4.6A). When comparing the same interaction between RING1B genotypes, I observed the same pattern to local compaction – perhaps, a slight reduction of

interactions in I53A cells, and complete disappearance of the distal interactions in cells lacking RING1B. After statistical testing of differential interaction frequency in the same way as above, I observe quantitative, but not statistically significant decrease in interaction frequency in I53A cells, and a highly significant loss in KO (Figure 4.6B). I investigated another example in the same way. I chose a region containing three prominent PRC1-associated loops, between *Nkx2-2*, *Pax1* and *Foxa2* (and a few less prominent loops which I didn't analyse), which visually had a very similar pattern, like all other looping interactions I observed: largely retained interaction frequency in I53A data, and complete loss in KO data (Figure 4.6C). And again I could statistically confirm my observations: all three of these loops were not significantly affected in I53A cells, but were lost (with varying level of significance) in cells lacking RING1B (Figure 4.6D).

I then went on to quantify these distal interactions genome-wide using *coolpup.py*. First of all, I wanted to confirm that this is indeed observed genome-wide, and that the interactions are fully lost in RING1B KO. To do that I performed genome-wide pile-ups of all intra-chromosomal interactions of CGIs across three conditions, since CGIs are the primary target of PRC binding in mESCs (Figure 4.7A). I analysed four different sets of CGIs: all CGIs, CGIs not bound by RING1B, CGIs bound by RING1B and CGIs both bound RING1B and marked by H3K27me3.

In the first two categories I observed close to no enrichment in all three conditions. When analysing RING1B-occupied CGIs, I observe high level of enrichment in both WT and I53A cells, which is reduced to almost the same low

level as in CGIs not bound by RING1B in KO cells. A similar pattern with even higher enrichment in the first two conditions is observed in the last category of CGIs. This is consistent with these interactions being a general property of at least a large fraction of RING1B-bound CGI, so that I still observe a high enrichment, and it is not lost during averaging of all the potential interaction sites.

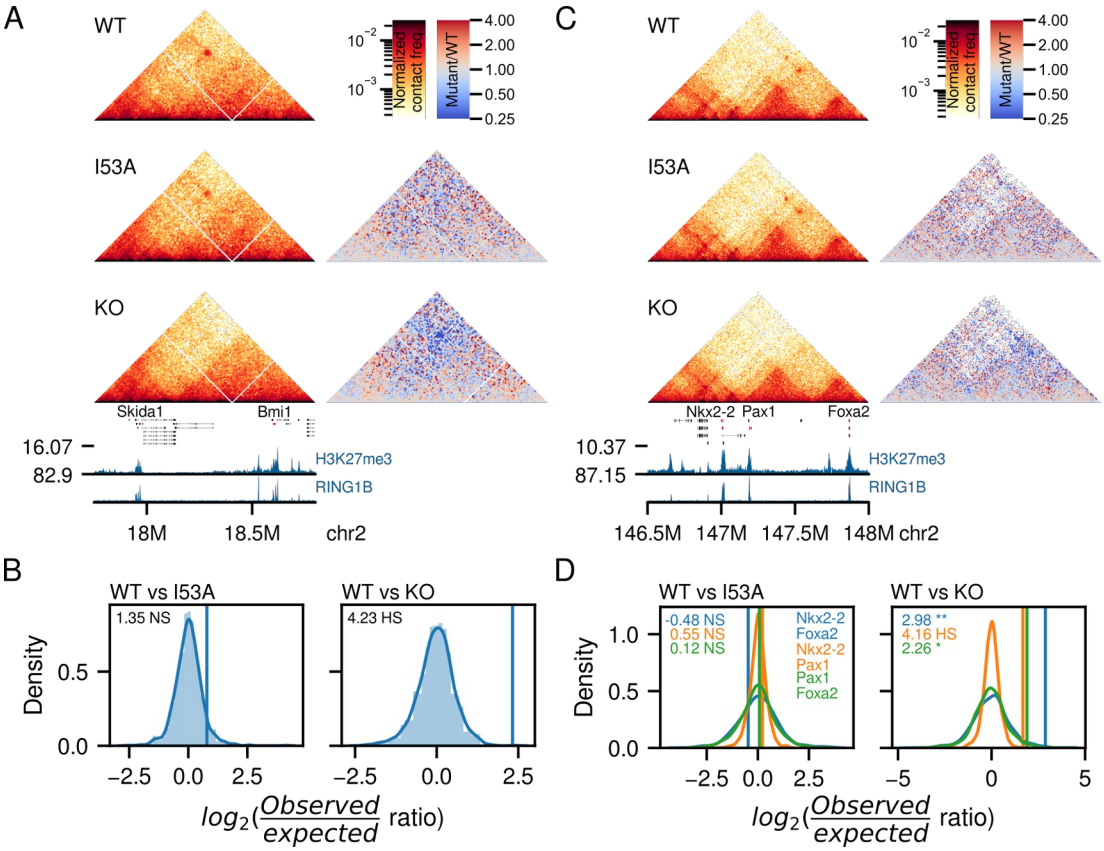

Figure 4.6. **Example of looping interactions between PRC1 targets. (A)** Same as Figure 4.3A for the region around a prominent interaction between PRC1 binding sites containing Skida1 and Bmi1 genes. **(B)** Same as Figure 4.3B, but for the Skida1-Bmi1 loop. **(C)** Same as (A), but for a region containing three pairwise interactions between PRC1 binding sites containing *Nkx2-2*, *Pax1* and *Foxa2* genes. **(D)** Same as (B), but for three interactions observed in (C). Colours identify which interaction is analysed.

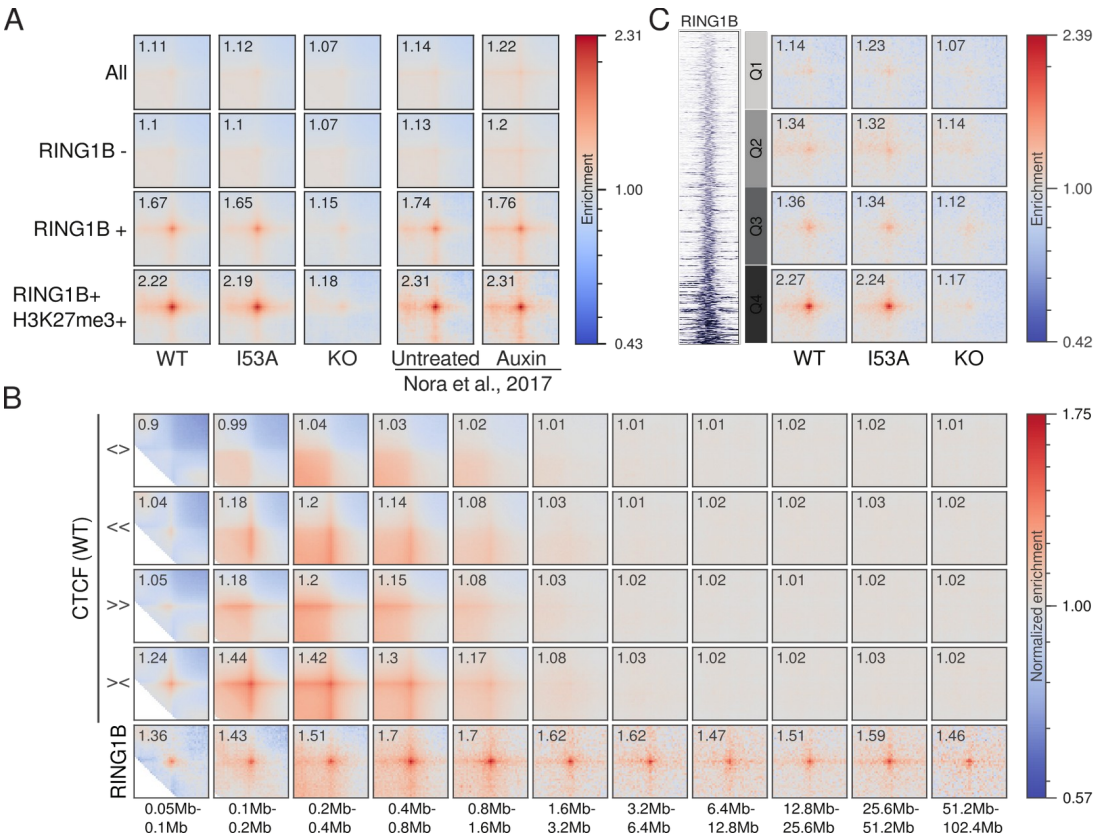

Figure 4.7. **Pile-up analysis of loops between PRC1 targets. (A)** Pile-ups of interactions between CGIs, normalized to expected level of interactions. In rows, subgroups of CGIs: all CGIs, CGIs with no RING1B binding, CGIs with RING1B peaks, CGIs with both RING1B and H3K27me3 peaks. In columns, different conditions: WT, I53A and KO cells, and then Untreated and auxin-treated CTCF-AID cells from (Nora et al., 2017). Text in top left corner shows enrichment in the centre of the pileup (average of central 3×3 pixels). **(B)** Pile-ups of interactions between CTCF sites of different orientation (> and <, for forward and reverse, respectively) and between RING1B peaks, at different distances in WT data. Here they are additionally normalized to average signal in top left and bottom right 3×3 pixels to avoid artificially high/low signal across the whole square at different distances. **(C)** Same as (A), except in rows are quartiles of RING1B peaks by total RING1B signal, and only for WT, I53A and KO data. Additionally, a heatmap of all ordered RING1B peaks is shown on the left (generated by Rob Illingworth).

To make sure these interactions are created by PRC1 rather than the well-

studied process of cohesin-driven loop extrusion (Fudenberg et al., 2016; Sanborn et al., 2015), I analysed published Hi-C data for ES cells in which CTCF can be efficiently degraded by addition of auxin (Nora et al., 2017). Strikingly, the pattern of average interaction enrichment for both untreated and auxin treated cells matches that of my WT and I53A (Figure 4.7A), although Hi-C experiments were performed in a different lab with a different protocol. This not only suggests robustness of my analysis to these variations, but also provides evidence for formation of PRC1-mediated interactions independently of loop extrusion. A recent report of enhanced RING1B-associated loops upon auxin-induced degradation of cohesin in mouse ES cells supports this conclusion, while also suggesting that loop extrusion can disrupt PRC1-mediated interactions (Rhodes et al., 2020).

From visual exploration of the Hi-C maps it was clear, that not all pairs of RING1B peaks form visible peaks of contact frequency enrichment. Therefore, I decided to investigate what rules govern formation of interactions between PRC1 binding sites. The first candidate was genomic separation between the peaks, since it has previously been shown that CTCF-mediated loops are restricted to a short distance range <2 Mbp (Rao et al., 2014). I hypothesized the same would be observed for RING1B-associated loops. To test this, I performed pileup analysis of CTCF- and PRC1-mediated interactions at different genomic distances (Figure 4.7B). I split CTCF peaks (from Bonev et al., 2017) by orientation of the CTCF motif in them (and discarded peaks with no motif, or motifs with different orientations, for simplicity). I then created all possible pairwise interactions of CTCF sites for all four combinations of

orientations (<>, >>, << and ><). Loops are predominantly formed by convergent (><) CTCF peaks; however, one can observe patterns formed by loop-extrusion for other orientations as well (Jung et al., 2017). I then performed pileup analysis using subsets of these interactions, and all potential interactions between RING1B peaks, across a range of increasing distances, starting from 50 kbp - 100 kbp with doubling separation intervals until 51.2 Mbp - 102.4 Mbp. All patterns in all four motif orientations for CTCF interactions disappeared at around 1.6 Mbp separation, but strikingly enrichment of looping between RING1B peaks persisted with similar strength across all distances. This means that distance is not a factor that influences interactions between PRC1 targets, and strongly suggests that the mechanism of their formation is distinct from loop extrusion, which creates CTCF-associated interactions. Of note, while the enrichment of interactions over background is clear at all distances, absolute interactions frequencies for these loops decay sharply with distance similarly to all intra-chromosomal interactions (Figure 4.7B).

To determine whether the local properties of the PRC1 targets play a role in determining the frequency of interactions between them I looked into whether simply the amount of PRC1 bound, as measured by RING1B signal in ChIP-seq data, has an impact on looping. I used RING1B ChIP-seq data to split peaks into quartiles by total read count, and then performed pileups for each quartile separately (Figure 4.7C). Interestingly, abundance of RING1B binding clearly impacts on distal interactions between PRC1 targets, since in particular Quartile 4 with highest binding had much higher enrichment of interactions than lower quartiles, in both WT and I53A data, while this pattern is

essentially absent from KO data. The weak increase in enrichment across quartiles in KO data is probably caused by a low level of RING1A-containing PRC1 in these cells. Taken together, these two findings suggest a mechanism for formation of PRC1-mediated interactions that is completely distinct from loop extrusion that creates CTCF-associated loops (consistent with a recent report (Rhodes et al., 2020)). These loops are most likely created by direct interactions between PRC complexes bound to the genomic sites, upon their (stochastic?) encounter in the 3D nuclear space. In this sense it is probably similar to the mechanism of driving nuclear compartmentalization based on interactions of like with like, and could involve phase separation (Plys et al., 2019; Tatavosian et al., 2018).

Any interactions involving simple search by diffusion and “stickiness” would be enhanced by longer interacting regions, since it would both facilitate more efficient exploration of 3D space, and promote stronger interactions between regions, once they have found each other. Therefore, I investigated whether the length of the RING1B peak regions impacts on interaction enrichment between them, using the same approach as with the RING1B analysis above: by splitting all RING1B peaks into quartiles by length, and performing pile-ups between regions in each group (Figure 4.8A). As predicted by my hypothesis, this showed a very prominent increase in interactions between longer RING1B peaks, similarly to analysis of total RING1B binding. However, of course, these two features are related to each other, since longer peaks will on average have more RING1B bound to them than shorter ones.

Another factor that could play a role in influencing interactions between RING1B peaks is the composition of the bound PRC1 complexes . It has been shown, on a few example regions, that canonical (PHC-containing) PRC1 (cPRC1) complexes can drive local interactions, while variant, or non-canonical, PRC1 (vPRC1) does not play a role in this (Kundu et al., 2017). That is consistent with the proposed structural role of cPRC1 components CBX2 and PHC, suggested from work both *in vitro* (Grau et al., 2011) and *in vivo* (Isono et al., 2013; Lau et al., 2017; Wani et al., 2016). Therefore I wanted to test this using genome-wide data using the same approach. I split RING1B peaks into 4 quartiles based on ratio of CBX2 and RYBP binding from ChIP-seq (Deaton et al., 2016; Rose et al., 2016)- canonical and variant subunits, respectively (Figure 4.8B). Interestingly, when sorting all RING1B peaks by this ratio, it produced a very prominent gradient in signal of both proteins, suggesting that different RING1B peaks are bound by different dominating subtypes of PRC1 complexes (Gao et al., 2012). Consistent with previous reports of structural role of cPRC1 subunits, Quartile 4 with high-CBX2/low-RYBP regions had a much higher interaction enrichment, than Quartile 1. The same result was obtained when using a different pair of cPRC1/vPRC1 subunits (MEL18, also known as PCGF2), and KDM2B (Blackledge et al., 2014; Morey et al., 2015); data not shown). This strongly suggests that cPRC1 is key for creation of distal interactions. However, the ratio of cPRC1/vPRC1 correlated with the amount of RING1B binding and peak length, which I have previously shown to be important factors for enhancing distal interactions.

Therefore, a new approach to analyse these relationships was required which would take all of the potential factors into account simultaneously.

I used the loop-ability of all RING1B peaks, defined as enrichment in the pile-up of each peak's interactions with all other RING1B peaks in the same chromosome (see previous Chapter). I then applied a linear model to predict the value of loop-ability using all potential factors discussed above. I also added H3K27me3 ChIP-seq (Illingworth et al., 2015) signal into the model, a mark deposited by PRC2, to make sure this approach works faithfully to recover only true factors important for looping, since PRC2 likely does not play a role in that (Illingworth, 2019; reviewed in Matheson and Elderkin, 2018). Since I normalized the range of values of the predictors in the models, their coefficients can be used as a measure of their quantitative impact on the outcome - loop-ability of a PRC1 target. A more positive value of the coefficient shows stronger positive impact on looping interactions, while negative coefficients suggest relatively negative influence of a predictor on interactions between PRC1 targets. I performed this modelling for WT, I53A and KO Hi-C data, and also for an independent wild-type ES cell dataset (Bonev et al., 2017). As expected, level of RING1B signal and Peak Length had a high positive impact on looping, while H3K27me3 had a negative effect (Figure 4.8C), in both WT and I53A data. A similar positive effect was observed for the components of cPRC1, in particular CBX2, while components of vPRC1 had a negative impact on looping. Interestingly, RING1B (and MEL18) occupancy are no longer statistically significant predictors of loop-ability, suggesting that while there are residual structures, they are not mediated by RING1B, but perhaps

by low level of RING1A-containing PRC1. As expected, the predictive power of the model dropped greatly for the KO data, as measured by Pearson's r between true and predicted values, since the interactions are much less prominent and variability in enrichment is probably mostly driven by noise. The WT result was validated by repeating the same analysis with a different wild-type ES cell dataset (Bonev et al., 2017). Since that dataset is much deeper, noise is reduced and both the coefficients of predictors and the correlation coefficient of the model are higher. Interestingly, the biggest increases in the model coefficient here relative to WT/I53A data correspond to cPRC1 components and Peak Length, consistent with these being the key properties important for looping.

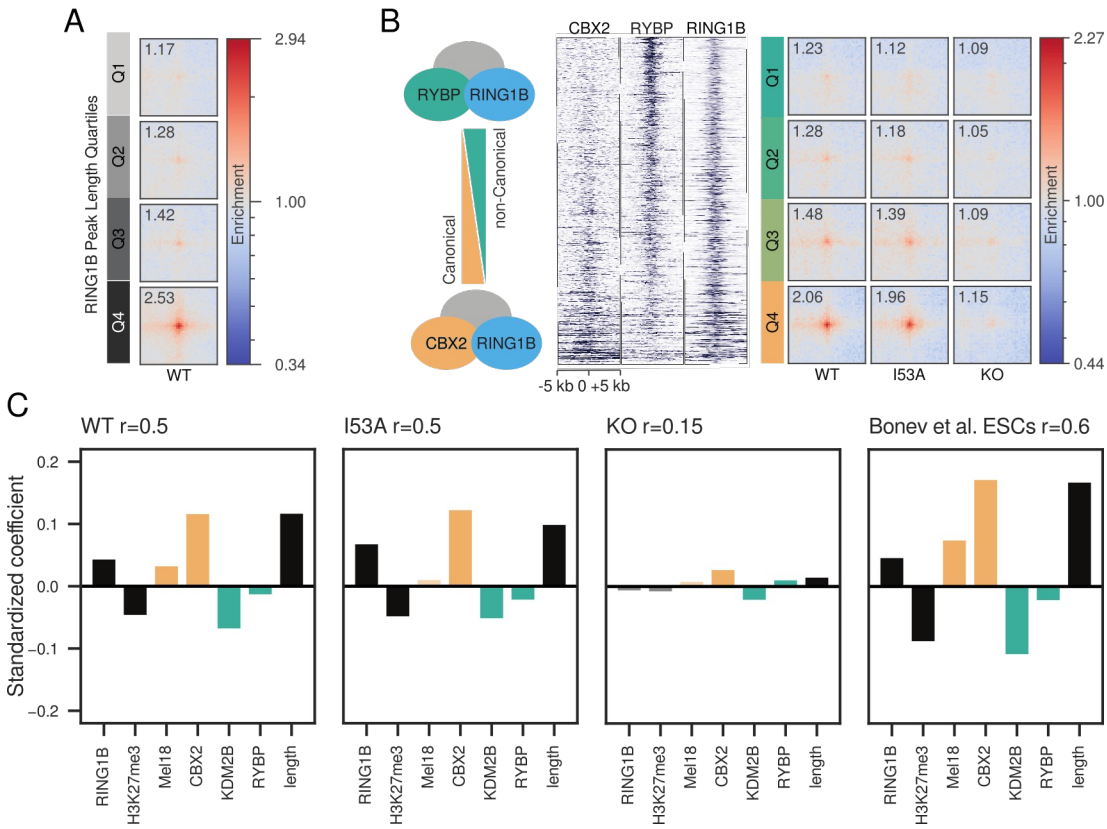

Figure 4.8. cPRC1 binding is key for establishment of distal interactions. (A) Pile-ups of RING1B peaks split into quartiles by their length in WT cells. **(B)** Pileups of RING1B peaks split into quartiles by ratio of CBX2/RYBP binding from ChIP-seq (in rows) for WT, I53A and KO cells (in columns). Heatmaps for CBX2, RYBP and RING1B ChIP-seq for all regions are shown on the left (created by Rob Illingworth). **(C)** Barplots showing coefficients for predictors in a linear model predicting loop-ability, in WT, I53A and KO cells, as well as an independent WT Hi-C from (Bonev et al., 2017). In orange, components of canonical PRC1, in green – variant PRC1. Lighter bars are not significantly different from 0 ($p>0.05$).

Interestingly, the RING1B peak regions that tend to loop the most to other peak regions, are similar to genes described in the literature as “retainers” of RING1B peaks upon loss of vPRC1 (Fursova et al., 2019) or some level of repression independently of RING1B catalytic activity (Blackledge et al., 2019): particularly long PRC1 targets with very high levels of RING1B binding.

Therefore I was interested in comparing these two sets. First, I needed to obtain the regions that I detect looping with others in my Hi-C data. I chose those that have loop-ability of at least 1.5 in WT cells, and lose it in KO cells by at least a factor of 1.5. Using this arbitrary definition, I obtained 295 RING1B peak regions that take part in looping interactions in a RING1B-dependent manner, that I termed “loopers” (Figure 4.9A). Loopers were significantly longer than non-loopers (Mann-Whitney p-value 5.38E-78), had more RING1B (Mann-Whitney p-value 1.36E-35) and CBX2 (Mann-Whitney p-value 8.87E-54) binding, consistently with my previous analysis (Figure 4.9B). While loopers also had higher enrichment for H3K27me3 (Mann-Whitney p-value 7.02E-13), they were not enriched for RYBP (Mann-Whitney p-value 0.057) and were depleted of KDM2B binding (Mann-Whitney p-value 8.4E-11) (Figure 4.9B). I then compared these regions to those, identified as “retainers” in previous studies (Blackledge et al., 2019; Fursova et al., 2019). I found the nearest RING1B peak regions corresponding to TSSs of the retainer genes, and intersected these sets with the set of loopers (Figure 4.9C).

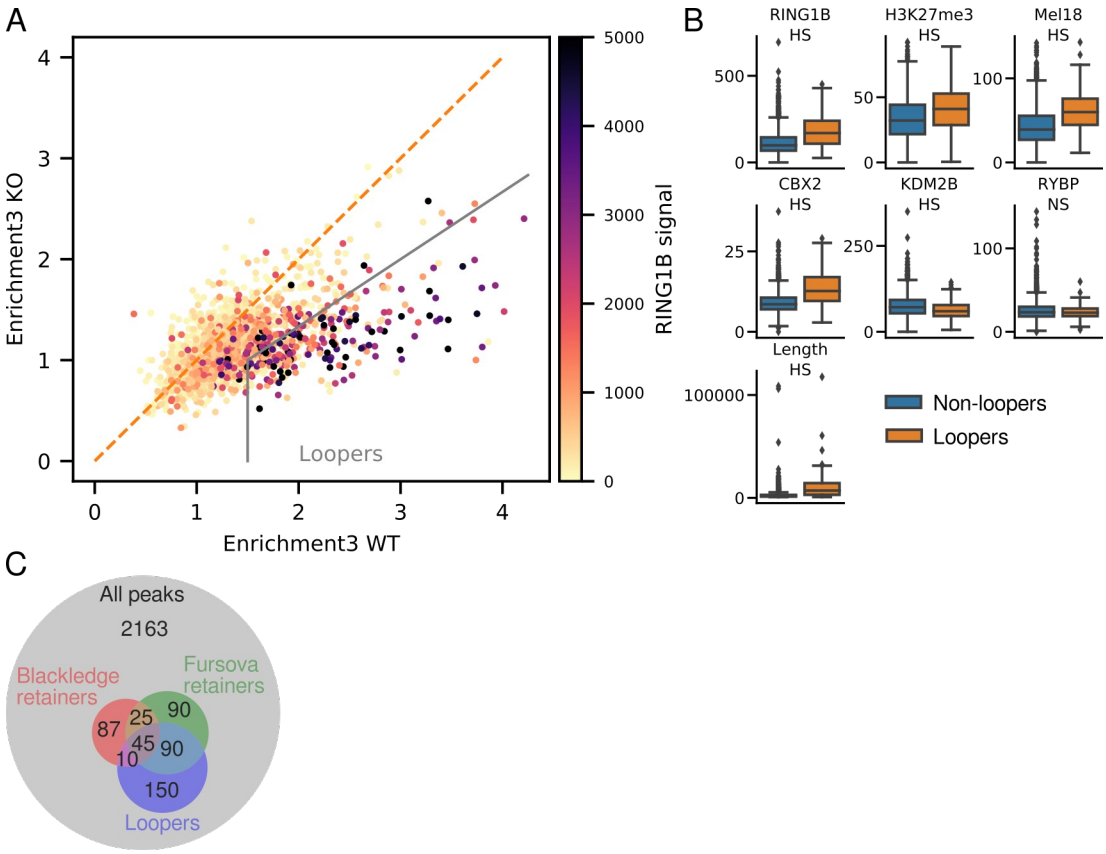

Figure 4.9. **Loopers retain repression in vPRC1 KO and RING1B catalytic dead cells. (A)** Scatter plot of loop-ability in WT (X axis) vs loop-ability in RING1B KO (Y axis) for all RING1B peak regions, colour-coded by total RING1B ChIP-seq signal. 295 regions with at least 1.5 loopability value and with at least 1.5× loss of interactions in RING1B KO cells are defined as "loopers". **(B)** Quantification of features of loopers and non-loopers: average signal of PRC1 components, H3K27me3 and region length. NS – Not significant, p>0.05, HS – highly significant, all $p<10^{-10}$.**(C)** Intersection of loopers and retainers from (Fursova et al., 2019) and (Blackledge et al., 2019). Both overlaps are highly significant using a hypergeometric test ($p=7.53\times10^{-74}$ and $p=2.83\times10^{-15}$, respectively).

Strikingly, more than half of the Fursova retainers are also loopers, a highly significant enrichment ($p=7.53\times10^{-74}$, hypergeometric test). Similarly, ~⅓ of Blackledge retainers are loopers ($p=2.83\times10^{-15}$, hypergeometric test) Considering the noisiness in the loop-ability measurements and the arbitrary

selection of loopers, there is a possibility these sets are even more similar in reality than what I observe. This suggests that these sites could be a special class of PRC1 targets, partially silenced by cPRC1, potentially through chromatin compaction, distal interactions and formation of polycomb bodies.

Finally, so far I have been assuming that loss of interactions in RING1B KO cells is due to absence of cPRC1 directly. One other possible explanation is that gene derepression and transcriptional activation are the reason for the loss of distal contacts. This would explain why the structures are retained in I53A cells, where changes to transcriptional landscape are minimal, but they are lost in RING1B KO cells with a large number of transcripts upregulated relative to wild type. To check this possibility, I again took advantage of the genome-wide nature of my data, and created pile-ups for RING1B peaks within different distances from strongly upregulated genes (Figure 4.10A). If gene activation is required for loss of interactions, I expected to obtain higher contact enrichment in the last category, that contains RING1B peaks farthest away from upregulated genes. However I didn't observe this. More importantly, I analysed RING1B peaks at varying separations from any upregulated genes (Figure 4.10B). Here, if gene expression is the cause of loss of interactions, I expected to see prominent loss of interactions only in the case of RING1B peaks close to an upregulated gene, but no difference for peaks far away. However I did not observe this, and if anything, the interaction enrichment is slightly higher for peaks closer to derepressed genes. These observations strongly argue against the hypothesis that transcription of target genes is what disrupts polycomb-

mediated looping interactions in RING1B KO cells, and, probably, loss of cPRC1 binding is the crucial factor.

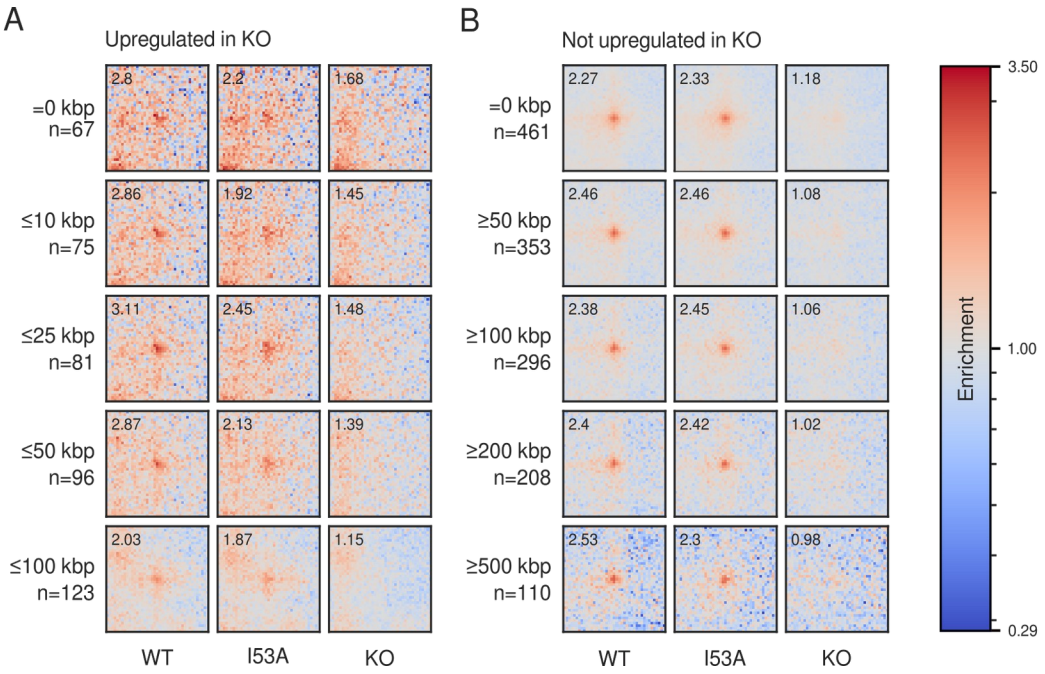

Figure 4.10. **Pile-ups of RING1B peaks at different distances from upregulated genes. (A)** Pile-ups between RING1B peaks from Quartile 4 (see above) at different distances from any genes strongly upregulated in KO over WT cells ($\log_2(FC) \geq 1$ & $p \leq 0.01$). Maximal distance in each category is shown on the left together with the number of such peaks in it; =0 kbp indicates overlapping peaks. WT, I53A and KO data in columns. **(B)** Same as (A), but for RING1B Quartile 4 peaks at least a certain distance away from any genes even weakly upregulated in KO over WT cells ($\log_2(FC) \geq 0.5$).

4.3 Discussion & conclusions

In this chapter, I have presented detailed analysis of 3D chromatin organization driven by PRC1 in mouse ES cells. I used Hi-C to investigate the structures associated with PRC1, after showing the high quality of the data. I used genome-wide averaging approaches to quantify changes in structures in RING1B I53A mutant and RING1B KO cells, while also discussing some example regions.

As has been shown previously, I observe local compaction of PRC1 targets genome-wide, as measured by Hi-C contact frequency. I show that this compaction is slightly impaired in I53A mutant cells, however only fully removing PRC1 by knocking out RING1B abrogates these structures completely. While this result is expected based on published FISH data, it contradicts a previously published paper from the Bickmore lab (Williamson et al., 2014). There, while observed by FISH, loss of compaction across the HoxD locus in RING1B KO cells was not observed by 5C, while knock-out of *Eed*, a core component of PRC2 showed an almost complete loss of compaction. A potential explanation for this discrepancy between two orthogonal methods is high abundance of PRC2 bound to HoxD in RING1B KO cells, which affects crosslinking, digestion or ligation efficiency during the preparation of 3C libraries. There are two likely explanations for the difference between 5C and my Hi-C data. First is the difference in the restriction enzyme that was used for digestion of the chromatin. In the 5C it was EcoRI, an enzyme blocked by CpG methylation (https://international.neb.com/products/r0101-ecori), while I used DpnII, a CpG methylation insensitive enzyme, when preparing Hi-C libraries.

Therefore, changes in DNA methylation around this region in different conditions could affect the 5C data, while Hi-C would not be affected. Another potential source of difference is data analysis. 5C data was not balanced to equalize visibility of all bins, and therefore a region with an unusually high total number of contacts also creates a visible increase in contact frequency within itself. Hi-C data are, however, balanced, and all locus-specific biases are removed.

RING1B catalytic activity has recently been shown to be crucial for PRC1/PRC2 targeting when inducibly and fully inactivated (Blackledge et al., 2019). I53A turns out to be a severe hypomorphic mutation, while it still causes an almost complete loss of H2AK119 ubiqitination levels. If this modification was directly important for creating chromatin compaction, I would expect to see a near-complete loss of structures associated with PRC1. However there is only a mild defect suggesting that it is mediated by somewhat impaired targeting of canonical PRC1 upon reduction of H2AK119ub.

This is also supported by absence of any effect of RING1B I53A mutation on looping interactions genome-wide, which are completely lost in RING1B KO cells. Why are local interactions more affected by level of H2AK119ub than distal interactions? The mechanism of their formation could be distinct from local compaction, and perhaps additional factors could be involved in compacting chromatin which interact with H2AK119ub, or therefore could be alternative substrates for RING1B E3 ligase activity; for example, cohesin-SA2 has recently been proposed to organize PRC1 domains in mouse ES cells (Cuadrado et al., 2019), unlike cohesin-SA1. This particular model is, however,

questionable in light of persistence of PRC1-associated structures upon auxin-inducible degradation of Scc1 (Rhodes et al., 2020).

My data suggest that PRC1 targets prefer to co-localize in the nucleus irrespective of the genomic separation between them (relative to random interactions at the same separations). This striking observation suggests a different mechanism for PRC1-mediated interactions from cohesin/CTCF-mediated loops extrusion, where interactions are restricted to 1-2 Mbp distance, and have no enrichment beyond that. While interactions between PRC1 targets at large distances have been observed previously (Bonev et al., 2017; Denholtz et al., 2013; Joshi et al., 2015; Schoenfelder et al., 2015; Vieux-Rochas et al., 2015), it hasn't been appreciated that the distance does not influence the interaction frequency enrichment between them. CTCF/cohesin mediated loops are generated by a linear process (loop extrusion, which acts along the chromatin fibre) and therefore are strongly influenced by genomic distance. cPRC1-mediated interactions, however, are probably not formed by a linear process: a simple "stickiness" of PRC1 targets upon stochastic coalescence in 3D space is sufficient to explain my observations (Figure 4.11).

Together with a clear role for cPRC1 subunits including CBX2, above suggests a (micro)phase/liquid-liquid phase separation mechanism of interactions between PRC1 targets, since CBX2 has be reported to be able to undergo this process (Plys et al., 2019; Tatavosian et al., 2018). Interestingly, unpublished observations of 1,6-hexanediol sensitivity of PRC1 target clusters made by Iain Williamson (a researcher in the Bickmore lab) support this hypothesis. An alternative mechanism of PRC1-mediated interactions involves oligomerization

of the PHC subunits through their SAM-domains (Isono et al., 2013). Since this interaction could link PRC1 complexes bound to distal targets, it would bridge genomic regions together. However such interactions, since they are specific protein-protein interactions and don't rely on weak hydrophobic interactions, should not be disrupted by 1,6-hexanediol. Possibly, PHC-driven interactions primarily act on shorter range local interactions and generate compaction, while phase separation driven by CBX2 is the key driver or distal contacts. Consistently with this, IW has preliminary data that suggest only partial decompaction of the extended HoxD region upon 1,6-hexanediol treatment, while distal regions behave similarly to RING1B KO cells. However definitive data regarding this hypothesis is currently lacking.

Finally, I showed that loss of gene repression is not the reason for the loss of distal loops in RING1B KO cells, since they are lost for interactions between RING1B binding sites not only near upregulated genes, but also those ≥500 kbp away from any genes with even slight increase in transcript levels. This also suggests that 3D organization of the chromatin by PRC1 is not sufficient to repress genes – its loss does not always lead to gene activation. This is consistent with a recent report of the importance of vPRC1 complexes in silencing the majority of PRC1 targets, with only a small fraction of genes silenced by cPRC1 (Fursova et al., 2019). This raises the question of the role for chromatin compaction and looping mediated by cPRC1. While in the flies these interactions have been shown to enhance silencing of PRC1 targets that take part in these interactions (Bantignies et al., 2003, 2011), this has not been shown for mammals, probably due to the difficulty caused by absence of PREs

in the system. Loss of the PHC subunits or of their ability to oligomerize causes loss of Polycomb bodies and leads to gene de-repression (Isono et al., 2013; Kundu et al., 2017), but it also might be causing reduction of PRC1 binding to its targets. Therefore it is unclear whether the 3D contacts and/or local compaction have a role in gene silencing.

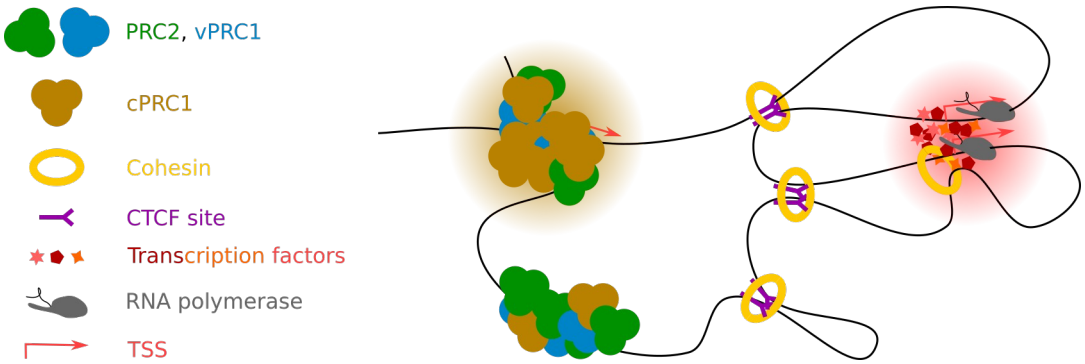

Figure 4.11. In addition to CTCF/cohesin-mediated loops and interactions between active genes, PcG-bound regions form long-range interactions via cPRC1. Both active regions and cPRC1 targets might interact due to phase separation of components of proteins in these genomic regions, and these interactions are not limited to short distances.

Chapter 5: 3D genome reorganization in ground state pluripotency

5.1 Introduction

Mouse ES cells are derived from inner cell mass of the E3.5 embryos. However, conventionally grown ES cells correspond to later developmental stages due to the presence of developmental cues in the foetal calf serum present in the media. An alternative approach for growing ES cells uses defined media without the presence of serum, but including two small-molecule inhibitors PD0325901 and CHIR99021, which target mitogen-activated protein kinase kinase (Mek) and glycogen synthase kinase-3 (Gsk3), respectively, and is termed “2i” (Ying et al., 2008). This culture method not only leads to uniformly high expression of pluripotency-associated factors such as Nanog, but also globally remodels the epigenome of ES cells: DNA is globally hypomethylated, which leads to redistribution of H3K27me3 and PRC binding (Joshi et al., 2015; Marks et al., 2012). While in serum-cultured ES cells the genome is coated with 5mC and it prevents polycomb-binding to non-CGI regions, in 2i H3K27me3 is extensively found in (no longer methylated) intergenic and repetitive regions (satellites), while its enrichment over CGIs is greatly diminished (Marks et al., 2012).

It has been shown by promoter capture Hi-C that certain polycomb-associated interactions are weakened in 2i culture conditions due to this epigenetic remodelling and redistribution of PRC binding (Joshi et al., 2015). I was interested in analysing this example of epigenetic remodelling in more detail genome-wide using Hi-C. I have shown that local compaction of extended PRC1 targets, such as Hox regions, is disrupted in 2i culture, and although not all regions respond to this PRC1 redistribution to a great extent, on average the

effect is easily observed. I then investigated interactions between distal PRC1 targets. This analysis revealed the same pattern: with overall reduced distal interactions, some regions lose long-range interactions completely, while at other loci the associations are only somewhat reduced. This seems to be related to the extent of RING1B binding loss in 2i relative to serum culture. Finally, I compared the average looping strength in serum and 2i culture to data from early mouse embryos. This revealed a striking absence of RING1B-associated loops in data from embryos, suggesting 2i models the epigenome and 3D genome architecture of ICM imperfectly.

The most important findings of this Chapter were included in our preprint (McLaughlin et al., 2019), which is currently under review in a peer-reviewed journal.

5.2 Results

5.2.1 Global analysis of Hi-C data from serum- and 2i-cultured mouse ES cells

I wanted to investigate reorganization of cPRC1-mediated interactions in a model of development that involves global epigenetic reprogramming of DNA methylation and polycomb distribution. In collaboration with Katy McLaughlin, a former PhD student in the lab, I performed Hi-C on ES cells grown in serum and 2i culture conditions. Cell culture and validation of 2i conversion were performed by Katy McLaughlin. Conversion to 2i was performed over 2 weeks. We generated 2 independent Hi-C datasets for each condition, to obtain a total of ~350 and ~400 million cis contacts longer than 1 kbp total for each condition (Figure 5.1A). Overall quality of the libraries was consistently high with >40% usable reads of all reads, with the only exception of serum-1, where it was ~28%.

I then analysed the curves of contact probability across distance separation (P_c(s)) for these data (Figure 5.1B). They were consistent between two replicates, but varied between conditions. In particular, both 2i libraries exhibited much higher contact probabilities at very long distances >3×10^7 bp. This is probably caused by changes in cell cycle in the ground state pluripotency, since contact probability at this distance range is much higher in G1 cells than other cell cycle stages (Nagano et al., 2017; Naumova et al., 2013), and cells cultured in 2i have a longer G1 phase (ter Huurne et al., 2017). I did not observe any other significant differences in the P_c(s) curves between culture conditions consistent with grossly similar chromatin organization.

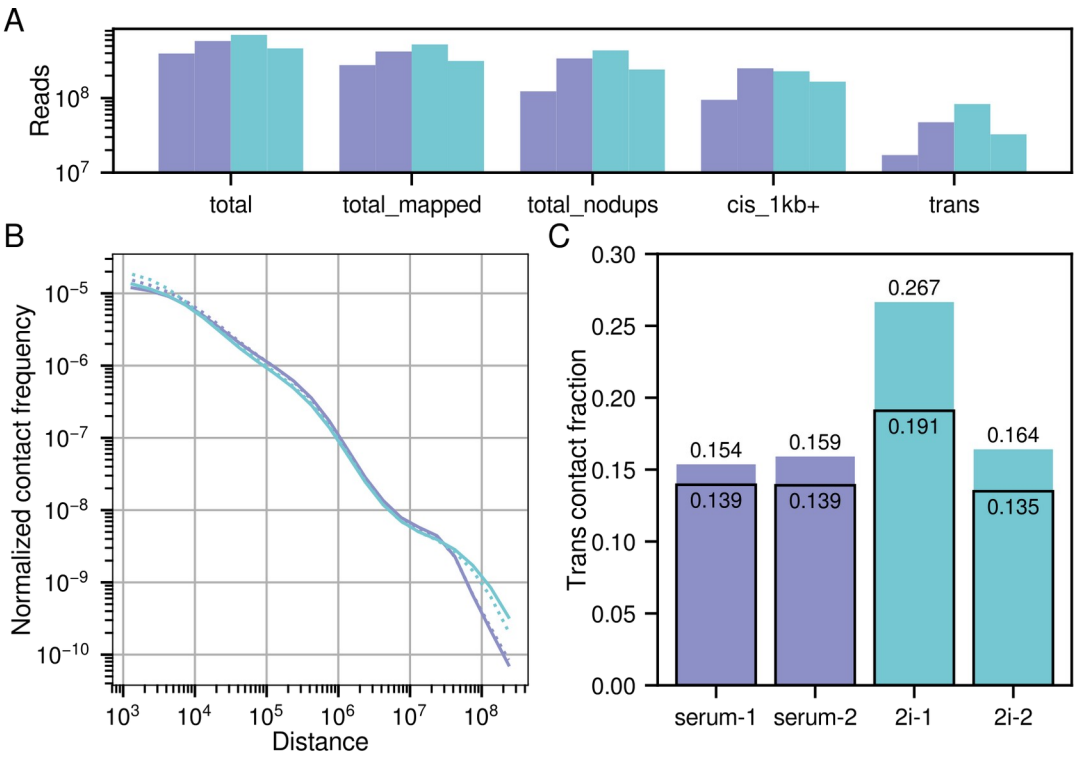

Figure 5.1. Quality control and comparison of gross properties of Hi-C data. **(A)** Read statistics of Hi-C datasets used in this Chapter: two replicates each of mouse ES cells grown grown in serum (purple) and 2i (light blue). Y axis shows number or reads, and is log-scaled. Categories are the same as in Figure 4.1 **(B)** $P_c(s)$ curves for all data used in this Chapter. **(C)** Fraction of inter-chromosomal contacts in all replicates of the data used in this Chapter. Bars show fraction of combined cis_1kb+ and trans contacts. Black boxes show fraction of total deduplicated contacts.

I then investigated the fraction of inter-chromosomal contacts found in these libraries (Figure 5.1C). While these were very consistent for two replicates of cell grown in serum, two 2i libraries had very different *trans*-contact fractions. Since I excluded very short-range *cis* contacts (<1 kb) when calculating these fractions, I suspected that perhaps the fraction of these short contacts is very different between the two replicates. Indeed, in 2i-1 35% of *cis* contacts are shorter than 1 kbp, and only 20% - in 2i-2. Therefore I repeated the trans-

contact fraction calculation without excluding short-range contacts, and observed much more similar values. However these values will now contain a lot of artefactual reads and their comparison between conditions is not meaningful. While this relies on a single replicate of data from 2i cells with lower fraction of <1 kbp contacts, the fraction of trans-contacts in ground state seems slightly elevated, which would suggest somewhat higher chromosome intermingling and lower level of chromosome compaction. Alternatively, this might also be related to relative enrichment of G1 cells in 2i.

Ground state ES cells have a significantly different gene expression profile compared to cells grown in serum (primed pluripotency) (Marks et al., 2012), therefore I wanted to check whether chromatin compartmentalization is also altered. I performed clustering analysis of the eigenvector tracks (Figure 5.2A), and also projected them onto the first two Principal Components (Figure 5.2B). The two culture conditions were clearly separated in both analyses. Similarly, I analysed local insulation profiles, which showed a separation of the two conditions, but with a smaller difference than compartmentalization (Figure 5.2C,D). Therefore while there are differences in insulation between ES cells grown in serum and 2i media, the most prominent global effect on 3D organization is related to compartmentalization, which is probably related to changes in gene expression or differences in cell cycle.

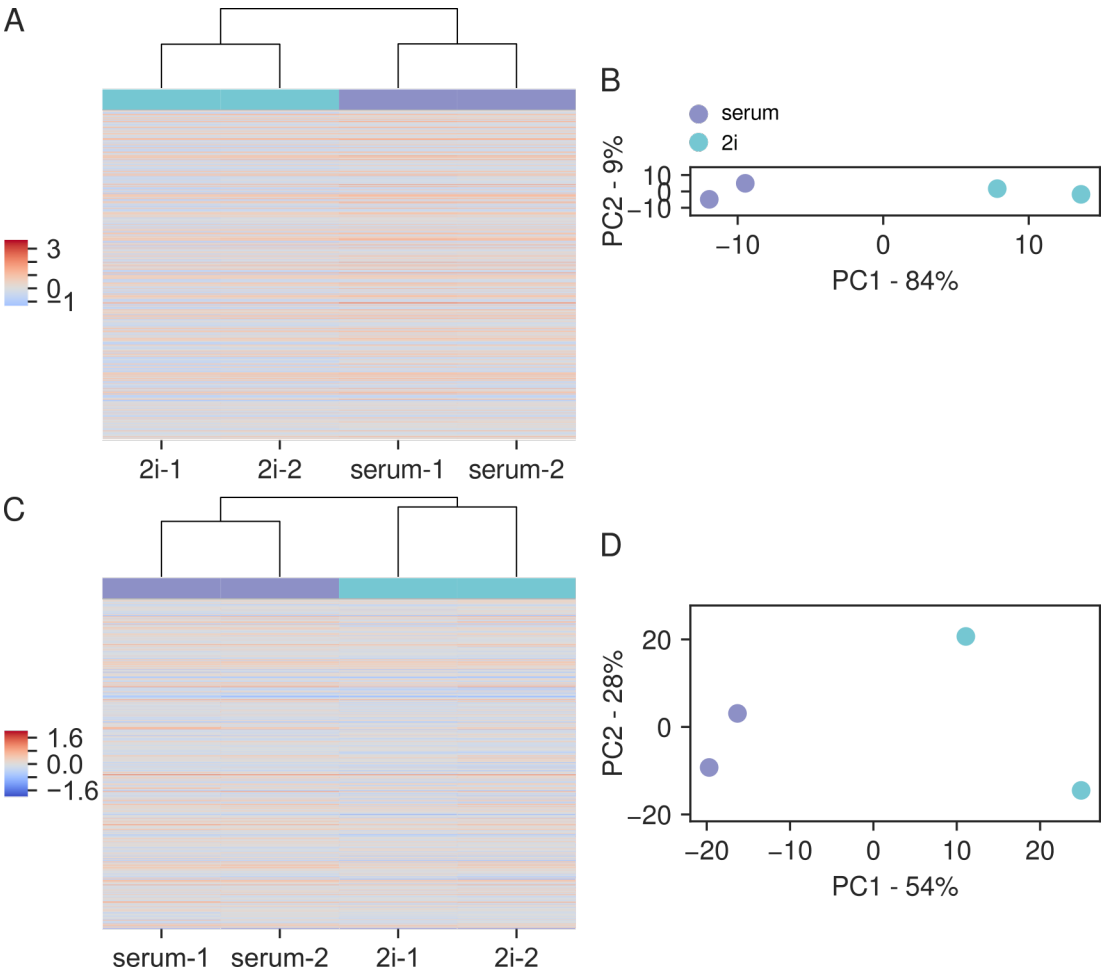

Figure 5.2. Comparison of compartmentalization and local insulation. (A) Hierarchical clustering of genome-wide compartment signal (first eigenvector) of Hi-C data from all Hi-C data used in this Chapter (200 kbp resolution). **(B)** Principle component analysis of data from (A). Colours of the points indicate culture conditions. **(C)** Same as A, but for insulation index (25 kbp resolution, 1 Mbp window size). **(D)** Principle component analysis of data from (C). Colours of the points indicate culture conditions.

5.2.2 Local compaction of extended PRC1 targets in ground state pluripotency

As discussed in the previous Chapter, long targets of PRC1 are strongly compacted in WT ES cells grown in standard serum/LIF culture conditions. Therefore, I decided to investigate whether this compaction is perturbed in 2i culture upon altered PRC1 binding due to global DNA hypomethylation. First, I investigated the 4 paralogous Hox loci, where polycomb binding over ~100kb domains has previously been shown to cause a visible chromatin compaction (Eskeland et al., 2010) (Figure 5.3). The most striking difference between the Hi-C data for these regions in serum-grown vs 2i cultures mESCs was observed at *HoxC* (Figure 5.3C). There is a loss of local interactions over HoxC in 2i conditions which, though not as complete as that in RING1B KO cells (see previous Chapter), it is clearly visible and highly significant with z=5.7, p-value=$5.9*10^{-9}$ (Table 3). Of note is almost complete loss of RING1B binding across the HoxC cluster in ground state (Figure 5.3C), which correlates with significant upregulation of transcription in the *HoxC12-HoxC13* region (Marks et al., 2012). *HoxA* and *HoxB* clusters also significantly lose interactions by Hi-C (Figure 5.3A,B; Table 3), while *HoxD* has a visibly reduced, but not significant, loss of interactions in 2i (Figure 5.3D, Table 3). This is perhaps due to the lowest level of transcriptional upregulation among all four Hox clusters, according to visual inspection of data from (Marks et al., 2012; McLaughlin et al., 2019). Even though changes in Hi-C contacts at *HoxD* were not significant, the *HoxD* locus is significantly decompacted in 2i conditions when measured by

FISH (McLaughlin et al., 2019). These experiments were performed by Katy McLaughlin.

When exploring non-Hox polycomb targets, a similar pattern is observed. While the *Cbx2/4/8* gene cluster visibly loses interactions (Figure 5.4A), this is not significant (Table 3), another similar region containing the *Nr2f2* gene is significantly decompacted (Figure 5.4B, Table 3). This suggests variability in the responses of different regions to DNA hypomethylation: while some loss of structure can usually be seen, only a subset of regions show statistically significant changes in Hi-C contacts.

To quantify the loss of local Hi-C interactions indicative of decompaction genome-wide, I used the same approaches as in the previous Chapter: I used local rescaled pile-ups of Hi-C data for all ≥10 kbp long RING1B target sites (n=181) and observed areas of enriched contact frequency corresponding to RING1B binding sites in serum-cultured ES cells in local pileups (Figure 5.5A). These were greatly depleted in cells grown in 2i culture conditions and show that, on average, genome-wide compaction of extended PRC1 targets is lost in ground state pluripotency. Importantly, it doesn't disappear completely, consistently with some regions retaining a certain level of compaction in this condition.

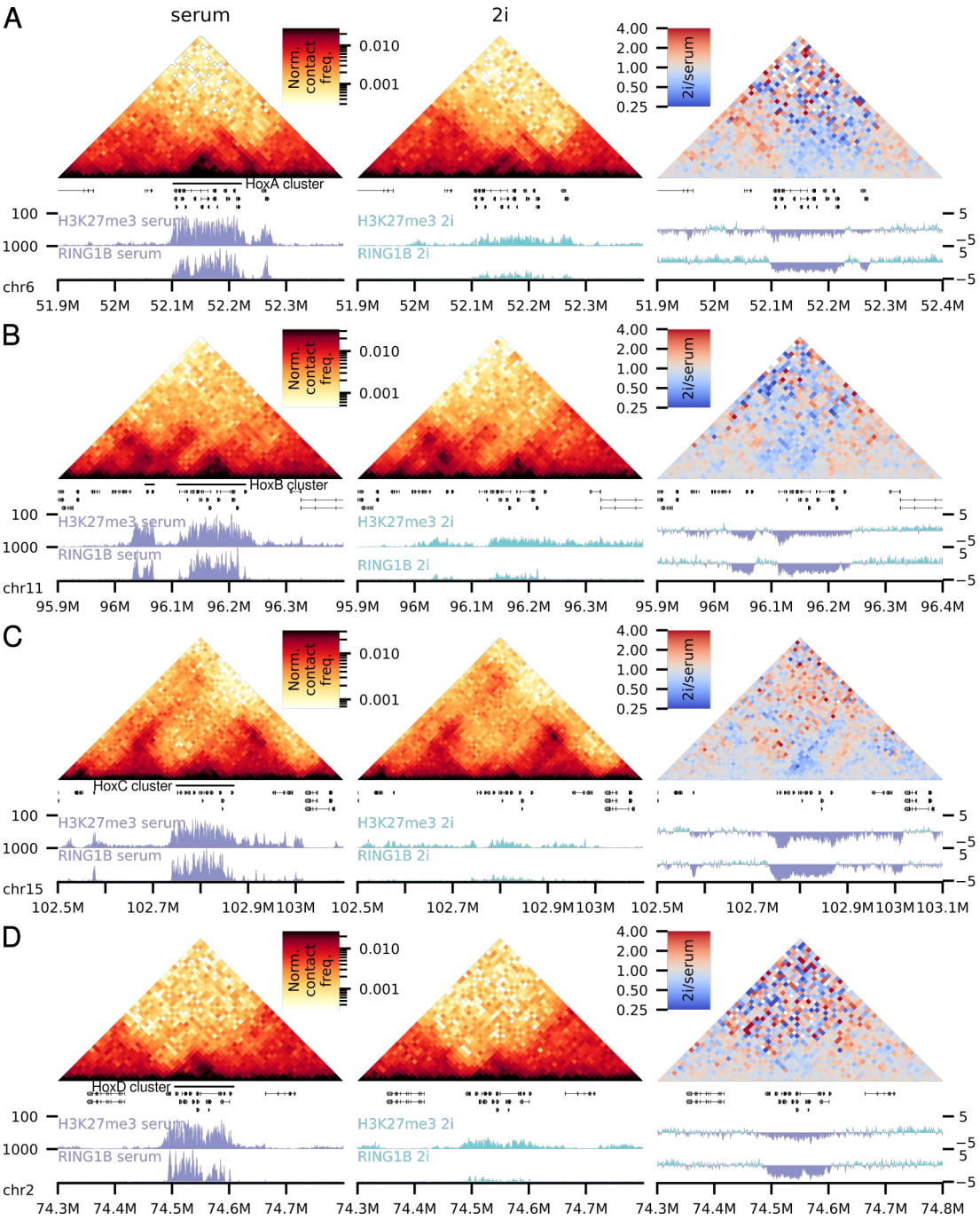

Figure 5.3. Local compaction of Hox clusters in serum and 2i culture. (A, B, C, D) Hi-C map of the (A) HoxA, (B) HoxB, (C) HoxC, (D) HoxD regions in serum and 2i, and a differential heatmap. Genes, H3K27me3 (Marks et al., 2012) and RING1B ChIP-seq (Joshi et al., 2015) are shown below, with $\log_2$(2i/serum) shown below the differential heatmap.

Region	Coordinates (mm9)	z-score	p-value
Cbx	chr11:118880000-118955000	1.21	0.114
HoxA	chr6:52100000-52220000	2.03	0.0210
HoxB	chr11:96110000-96215000	2.94	0.00167
HoxC	chr15:102740000-102870000	5.70	5.87E-09
HoxD	chr2:74495000-74605000	1.09	0.137
Nr2f2	chr7:77425000-77520000	2.47	0.00672

Table 3: statistical analysis of change in local compaction between serum and 2i for regions discussed in this Chapter: four Hox regions, the Cbx gene cluster and the *Nr2f2* locus. P-values<0.05 are highlighted in green.

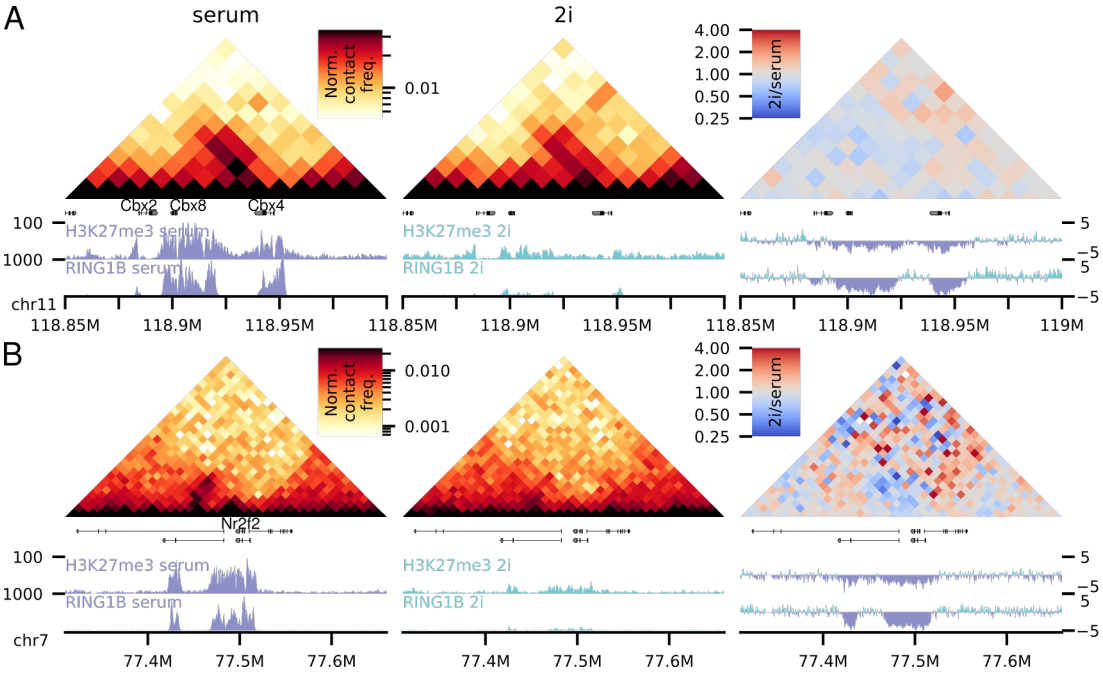

Figure 5.4. Local compaction of non-Hox regions in serum and 2i culture. (A) Same as Figure 5.3A, but for the *Cbx2/8/4* gene cluster. **(B)** Same as (A), but for the *Nr2f2* gene region.

I also plotted the average number of observed/expected Hi-C contacts in 25 kb genomic windows split into quantiles by the level of RING1B occupancy (Figure 5.5B). A significant interaction enrichment in the highest RING1B group is observed in serum-cultured cells, with a clear enrichment in the second-highest group too. Interestingly, unlike the WT cells from the previous chapter the

highest RING1B group has even higher contact frequency than the no-RING1B group; perhaps, this is related to slightly different medium used when growing these cells with 15% FCS instead of 10%. While there is a slight increase in contact frequency in the top two groups in 2i cells relative to the low-polycomb group, it is significantly lower than in serum, consistently with the previous analysis. This showed that loss of Hi-C interactions in 2i conditions is restricted to the ~0.1% of the genome with the very highest levels of RING1B occupancy. An important note required here is that the absolute levels of RING1B in each category is different between serum and 2i cells: the range of RING1B signal in ChIP-seq is much smaller for 2i, and therefore the same quantiles correspond to very different protein occupancy. However more direct comparison is impossible as, without normalisation to a spike-in control, ChIP-seq is not quantitative. The dip observed in the highest RING1B in one of the 2i replicates (2i-1) is probably an artefact due to the high frequency of short-range contacts (see above). Repeating this analysis using H3K27me3 instead of RING1B ChIP-seq data gives the same results (data not shown).

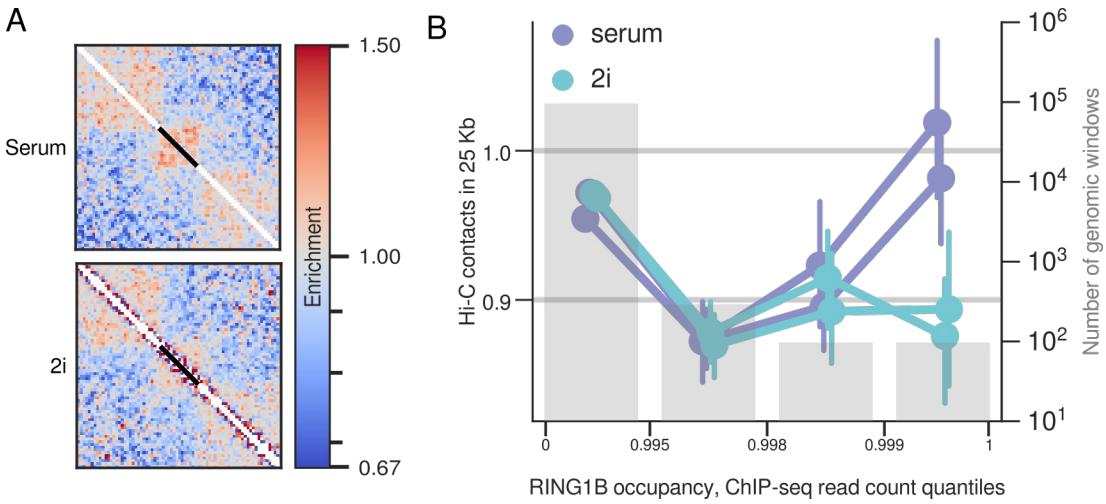

Figure 5.5. **Genome-wide quantification of local compaction of PRC1 targets in serum and 2i culture conditions. (A)** Local rescaled pileups of all long regions of RING1B binding (length>10 kbp, n=181) in serum and 2i Hi-C data. Black bar shows the location of the averaged RING1B binding sites. **(B)** Mean number of normalized local Hi-C interactions in 25 kbp windows of varying levels of RING1B binding from ChIP-seq ((Joshi et al., 2015), split into four groups), for serum and 2i data. Error bars show 95% confidence interval obtained by bootstrapping. Shown separately are curves for both replicates. Purple, serum; light blue, 2i. Number of regions in each category is shown as grey bars with value on the right Y axis.

5.2.3 Interactions between distal PRC1 binding sites in ground state pluripotency

As discussed in the previous Chapter, PRC1 target sites coalesce in the 3D nuclear space. Therefore, I wondered how these interactions would be affected by the polycomb redistribution and its loss from many of its normal targets in 2i cultured ESCs. First, I visually and statistically analysed some examples of PRC1-associated loops found in my Hi-C data. The *Skida1-Bmi1* loop, which I also focused on in the previous Chapter, is prominently observed in the serum

data, but completely disappears in 2i (Figure 5.6A). This is aligned with very efficient loss of RING1B binding across the *Skida1* locus (Joshi et al., 2015). This reduction in interaction frequency is highly statistically significant with z-score of 4.2 (Figure 5.6B), similar to that observed in RING1B KO data. I then investigated other example loci. Here I show the region of chromosome 5 containing major RING1B peaks over *En2*, *Shh* and *Mnx1* (Figure 5.6C). While the interactions between them are less obvious than those at *Skida1-Bmi1* due to shorter RING1B peak length, they are all quantitatively reduced in ground state ES cells, with the *En2-Shh* loop showing statistically significantly decreased contact frequency (z-score of 2.28) (Figure 5.6D). The much weaker loss observed here can be explained by two factors: first, lower original enrichment in serum-cultured cells; second, low level of residual PRC1 binding in 2i culture, compared to complete loss at the *Skida1* locus.

I then decided to investigate the loss of looping between PRC1 targets genome-wide. First, I confirmed it can be observed by quantifying looping between CGIs in the two culture conditions (Figure 5.7A). I did not observe a significant enrichment of interactions between RING1B-negative CGIs in either condition. However, as expected, RING1B-positive CGIs interacted very prominently with each other in serum cultured cells, while the enrichment was reduced in ground state pluripotency. The partial loss of interactions is consistent with the earlier examples, where certain regions lose interactions completely, and others only partially.

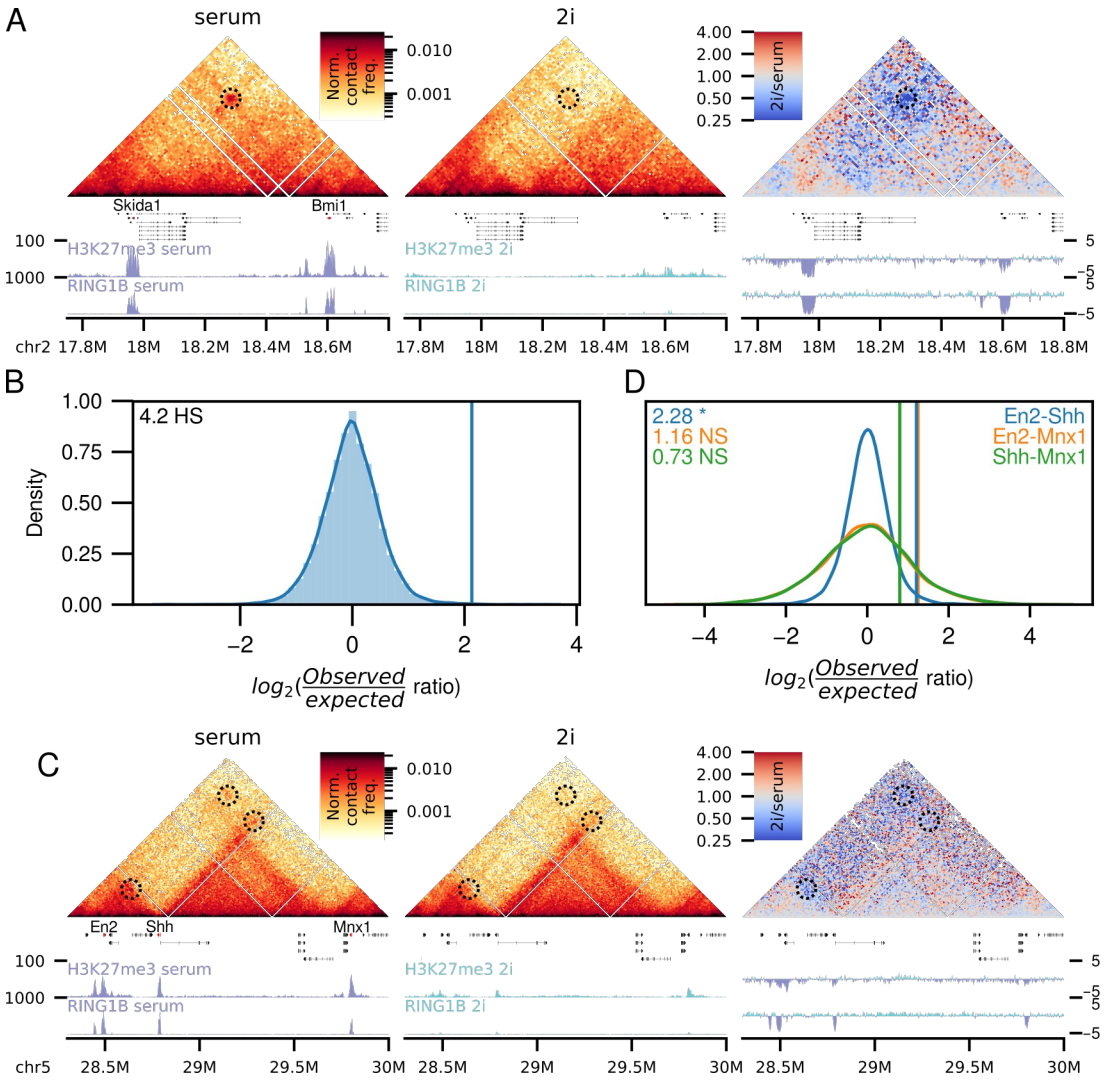

Figure 5.6. Example of looping interactions between PRC1 targets in serum and ground state. (A) Same as Figure 5.3A, for the region around a prominent interaction between PRC1 binding sites containing Skida1 and Bmi1 genes. **(B)** Same as Figure 5.3B, but for the Skida1-Bmi1 loop. **(C)** Same as (A), but for a region containing three pairwise interactions between PRC1 binding sites containing En2, Shh and Mnx1 genes. **(D)** Same as (B), but for three interactions observed in (C). Colours identify which interaction is analysed.

I then asked which regions tend to lose Hi-C interactions the most in 2i. I hypothesized, in line with my visual exploration of the data, that the most prominent RING1B peaks would show the biggest loss of RING1B binding, and

therefore the biggest loss of interactions with other PRC1 targets. To analyse this quantitatively, I performed loop-ability analysis of interactions between each RING1B peak and the rest of the peaks on the same chromosome (see Chapter 2). After doing this for both serum and 2i data, I compared the obtained enrichment values with the RING1B ChIP-seq signal within the peaks in these two conditions (Figure 5.7B). I observed a clear and highly significant reduction in loop-ability in 2i conditions relative to serum culture (Wilcoxon $p=2.1\times10^{-88}$), consistent with the general loss of interactions between PRC1 targets. Moreover, both by visual analysis and using linear regression, I observed that the loss of interactions was highest for those regions, that highly interact in serum media, while those with low loop-ability were not affected to the same extent. Consistently with this, the highest loop-ability regions tended to have very high levels of RING1B binding in serum, and they were severely reduced in 2i culture, while those with higher preservation of loop-ability had higher levels of residual PRC1 in 2i. Unfortunately exactly quantifying these patterns is impossible due to non-quantitative nature of ChIP-seq, dependency of loop-ability on the chromosomal context of the region and potential non-linear relationships between these different kinds of data.

Mouse ES cells grown in 2i are in ground-state pluripotency, and resemble the *in vivo* population of the inner cell mass (ICM) of E3.5 embryos (Habibi et al., 2013; Marks et al., 2012; Wray et al., 2010). Therefore, I was interested in comparing 3D chromatin structure in 2i-cultured cells with that of the ICM. I found 3 publicly available Hi-C dataset for ICM or whole E3.5 embryos (Du et al., 2017; Ke et al., 2017; Zhang et al., 2018) which I used in this analysis. I

then performed pile-up analysis of loops, identified in ES cells and split into CTCF- and RING1B-associated groups (Figure 5.7C), since this approach is more sensitive than analysis of local compaction, and depends less on data quality. I used my Hi-C data from serum and 2i cells together with the published Hi-C data from embryos. One of the papers also generated Hi-C from (serum-cultured) ES cells using the same protocol, so I used that data too to ensure their method generates libraries of sufficient quality. This analysis showed a big reduction in RING1B-associated loop strength in 2i-cultured cells relative to serum culture, while CTCF-associated loops seem slightly enhanced. All other data have lower enrichment in CTCF-associated loops, including the mESC data (Du et al., 2017) (probably due to deeper sequencing and/or higher quality of our data), however it is similar across datasets. Similarly, RING1B-associated loops appear slightly weaker in the published mESC data than in our data from serum-cultured cells, but still clearly different from the weak enrichment of 2i-grown cells. In contrast, all three datasets of Hi-C from E3.5 embryos (or ICM cells) show no enrichment for RING1B-associated loops. This suggests complete lack of RING1B-associated looping *in vivo* at this stage. Therefore, while in ground state cells the 3D genome is reorganized in the direction towards what I observed *in vivo*, these loops are still clearly present in 2i culture. When analysing data from later embryos (E6.5, E7.5) (Zhang et al., 2018), I observed a high enrichment at sites of RING1B-associated loops, consistent with much higher levels of DNA methylation at these stages.

It might seem like a drawback of this analysis is the potentially different distributions of PRC1 binding *in vivo* and *ex vivo* in ES cells. And while this

indeed can be the reason for observed lack of looping between RING1B targets in the early embryo, it would still mean the epigenome of ICM is very different from that in 2i-cultured ES cells, and they are not a faithful representation of the true embryonic pluripotent state, while probably more similar than the serum-cultured cells.

Above I mentioned higher enrichment of CTCF-associated loops in 2i Hi-C data. Since CTCF and SMC1a ChIP-seq data for serum- and 2i-grown mESCs recently became available (Atlasi et al., 2019), I decided to check if their binding is enhanced in the latter condition, I re-analysed the published ChIP-seq data. I created average profiles and heatmaps for CTCF sites taken from serum-cultured cells from (Bonev et al., 2017) (Figure 5.8A, B) and observed higher enrichment of ChIP signal for both CTCF and SMC1α in 2i data for both, motif orientations. Additional “shoulders” of enrichment are observed in data from 2i-cultured cells, and the shoulder downstream of the CTCF sites (“inside the loop”) is higher, subtly for CTCF and very clearly for SMC1α. This suggested more CTCF and cohesin binding in the vicinity of serum-defined peaks, with a preference for the downstream direction; this would additionally enhance the boundary strength in 2i that causes increased enrichment of CTCF-mediated loops.

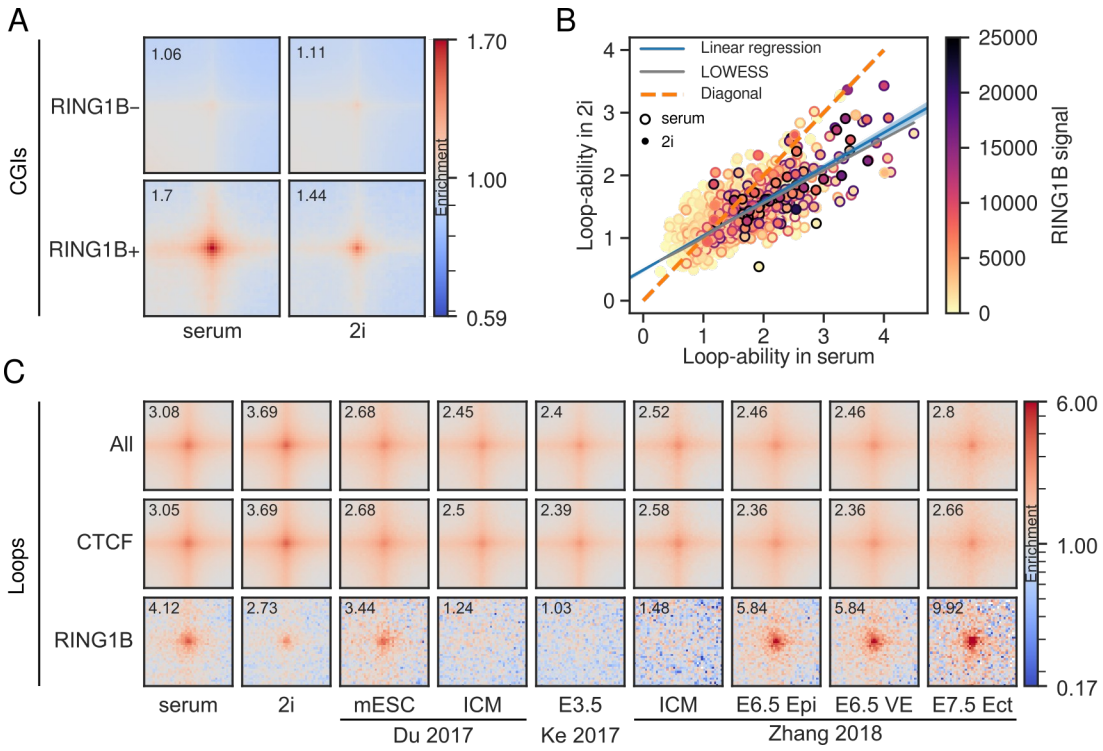

Figure 5.7. Genome-wide loss of looping between PRC1 targets in 2i culture. (A) Pile-ups of interactions between CGIs. In rows, subgroups of CGIs: CGIs with no RING1B binding, CGIs with RING1B peaks. In columns, different conditions: serum-grown and 2i-grown cells. Text in top left corner shows enrichment in the centre of the pileup. **(B)** Comparison of the loop-ability enrichment values of all RING1B peak regions between serum and 2i (X and Y axes), and colour-coded RING1B abundance in those peaks by ChIP-seq (Joshi et al., 2015): circles correspond to 2i, while rings around them correspond to serum. Dashed orange line shows the diagonal of equality, grey line is showing the LOWESS fitting curve, and the blue line – a linear model (with shaded 95% confidence interval obtained by bootstrapping). **(C)** Pile-ups of loops (in rows: all, CTCF-associated and RING1B-associated) across conditions (in columns: serum and 2i-cultured cells, ES and ICM cells from (Du et al., 2017), E3.5 embryos from (Ke et al., 2017), and ICM, E6.5 epiblast (Epi), E6.5 visceral endoderm (VE), and E7.5 ectoderm (Ect) from (Zhang et al., 2018).

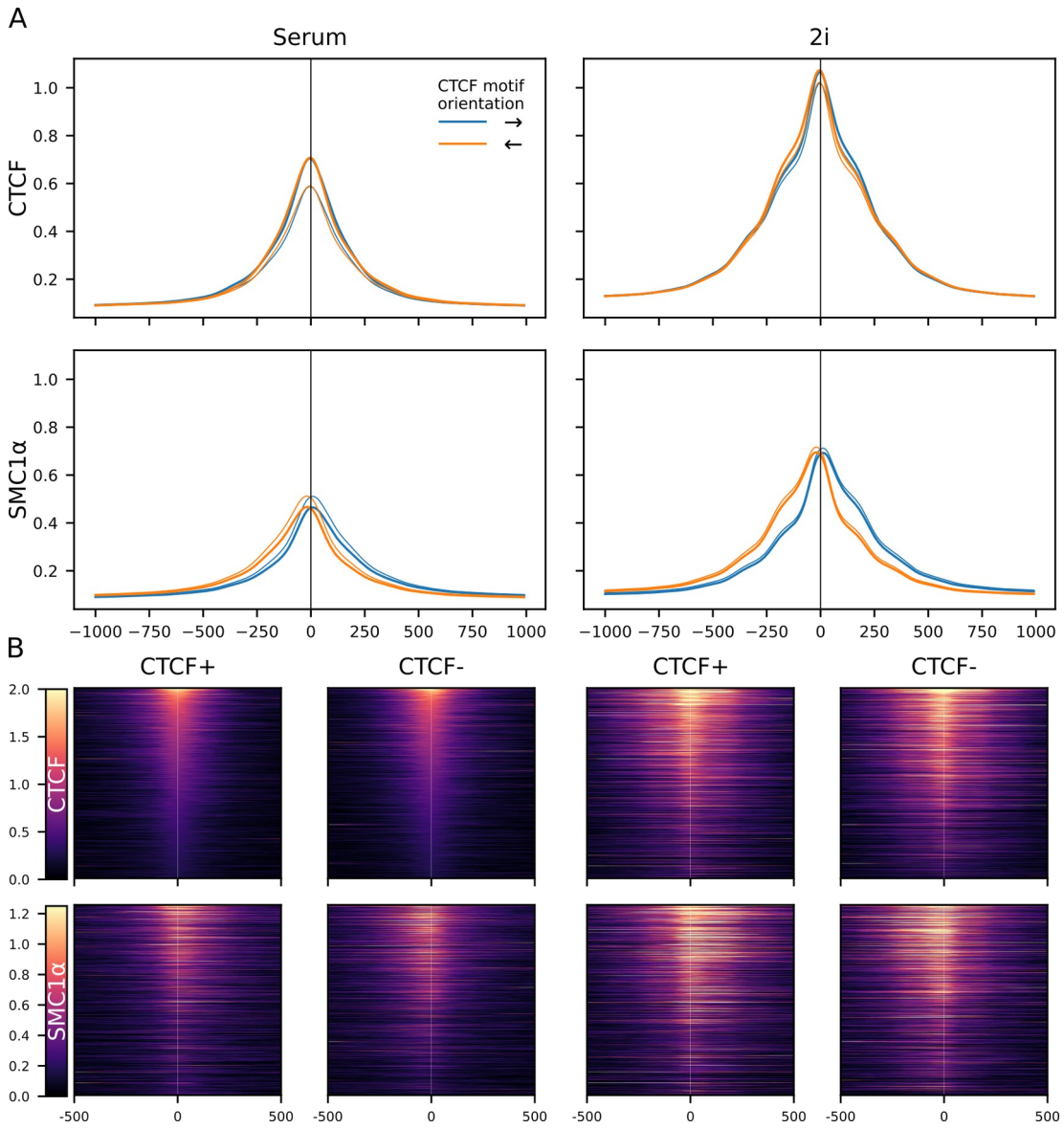

Figure 5.8. CTCF and SMC1α binding at CTCF peaks in serum- and 2i-cultured mESCs. (A) Average profiles of CTCF and SMC1α ChIP-seq for the top 20,000 CTCF peaks in serum and 2i-cultured cells. Peaks split by CTCF motif orientation. Two replicates of each dataset are shown, distance (bp) relative to the centre of the peak along the x axis, average ChIP-seq signal on y axis. **(B)** Heatmaps for the zoomed in data in (A); only replicate 1 is shown.

5.3 Discussion & conclusions

In this chapter, I have presented analysis of 3D chromatin re-organization in mouse ES cells grown in 2i relative to serum. I used Hi-C data from ES cells grown in these two conditions, and showed loss of local compaction in a subset of extended targets of PRC1, with a big reduction in contact frequency on average genome-wide. Similarly, interactions between distal PRC1 binding sites were reduced in 2i cultured cells on average, with some loops lost completely and others only slightly weakened. The variability is probably simply due to varying levels of loss of PRC1 at different genomic sites, however why some regions lose PRC1 binding more than others is an interesting question outside the scope of this work.

It is interesting that such drastic changes in the epigenome of ES cells, as global DNA demethylation and the following redistribution of PRC1/2 binding away from their “normal” targets (CGIs), is only correlated with minor changes in gene expression (Marks et al., 2012). Moreover, the majority of genes whose promoters lose H3K27me3, actually have lower expression in 2i (Marks et al., 2012). This suggests that simple (incomplete) loss of PRC1/2 is not sufficient to drive derepression of most genes, and specific transcription factor binding is required. A tempting hypothesis is that the loss of 3D structures would correlate with changes in transcriptional state of the genes. Although this does not seem to be the case based on examples from published capture Hi-C analysis (Joshi et al., 2015), a more comprehensive analysis is required to be sure gene upregulation and loss of PRC1-associated 3D organization are not linked in this system.

An important question is whether this epigenetic remodelling and 3D genome reorganization is important for establishment of the ground state. While not addressed in this Chapter, we answered this question in our manuscript (McLaughlin et al., 2019). Overexpression of DNMT3A or DNMT3B together with DNMT3L from an exogenous promoter in *Dnmt3a/3b* knock-out ES cells, done by Richard Meehan's group, leads to constitutively high level of DNA methylation even in 2i culture conditions. This ensures that polycomb binding is largely unchanged in ground state and leads to serum-like epigenome in 2i culture conditions. Katy McLaughlin used FISH to show that PRC1-mediated 3D chromatin organization is also preserved in DNMT3A/3B +3L expressing ESCs under 2i conditions. But strikingly, the transcriptional profile of these cells is very similar to that of wild-type ES cells grown in 2i. Culture conditions, and the signalling pathways affected by them, are therefore key in establishing the ground state pluripotency, while the changes in the epigenome and associated 3D genome are largely dispensable.

2i and serum media for ES cells model distinct developmental stages of the early embryo, at least according to their epigenome: low level of DNA methylation in 2i correspond to that observed in the inner cell mass of E3.5 embryos (Ficz et al., 2013; Leitch et al., 2013; Marks et al., 2012; Ying et al., 2008), while serum-cultured cells resemble a slightly later post-implantation stage (Ficz et al., 2013). An important distinction between *in vivo* developmental pathway and serum-2i transition, is the direction. In the embryo cells rapidly lose methylation from the zygote stage during cleavage divisions, until the E3.5 embryo, which corresponds to the 2i cultured cells. And then the

DNA methylation levels rise, and reach those similar to that in ES cells cultured in serum at E6.5 (Ficz et al., 2013; Zhang et al., 2018). However, during serum-2i conversion the cells start with a high level of DNA methylation, which is then lost. This might be responsible for the differences between 2i and ICM observed here. Possibly, establishment of robust polycomb domains in serum while DNA methylation levels are high, allows for their partial preservation in 2i: while PRC1/2 are partially drawn away from CGIs to the unmethylated genome, local polycomb signalling still maintains some level of binding. This could be investigated by analyzing ES cell lines established from the embryo using 2i media, and by transferring them into serum culture.

After E3.5, levels of DNA methylation increase, and PRC1-associated loops suddenly appear very prominently at E6.5 in both epiblast cells and visceral endoderm, and they are even stronger in E7.5 ectoderm. This is consistent with the model of PRC1/2 starting to bind specifically to CGIs upon increase in DNA methylation level. Interestingly, however, another potential reason for this dynamic is the changes in levels of PRC1/2 subunits. While, to my knowledge, there is no information regarding that for such early developmental stages, it has been reported that levels of some of the components increase during early development which would be consistent with absence of interactions in the early embryo, however stages beyond blastocyst were not analysed (Liu et al., 2016). Measuring global levels and binding profiles for PRC1/2 subunits during early development, together with more Hi-C datasets, would be very useful for more in depth analysis of 3D chromatin structure dynamics.

Chapter 6: Discussion

During my PhD, (i) I developed a tool to perform pileup analysis of Hi-C data, *coolpup.py*, and showed its usefulness in addressing biological questions using published data, (ii) I investigated the role of PRC1 in organizing the genome of mouse ES cells in 3D space and showed that catalytic activity of RING1B is not directly involved in creation of long-range interactions and local compaction, however, I showed that the presence of canonical PRC1 complexes is key to create interactions, and (iii) I investigated reorganization of the 3D genome in ES cells upon conversion to ground state pluripotency, which involves global changes in Polycomb binding, and I showed the loss of PRC1-associated structures in this condition. While my results were discussed in each chapter, here I would like to consider future directions in relation to each results Chapter, in turn.

Advancements in Hi-C data analysis require, first, finding optimal solutions for standard questions. In particular, the unification of the storage format would be highly advantageous for the field: then comparing different algorithms and datasets directly would be feasible without full reprocessing the data. Currently, two main formats exist: *.hic* and *.cool*. Both are efficient binary formats that store sparse representation of the data. However, they differ in some key properties. First, *.cool* files are based on the HDF5 storage format - commonly used by the wider scientific community, which ensures the information in these files can be accessed even if the whole current scientific ecosystem changes and the *cooler* CLI or the python library are no longer available. A related benefit is that writing a library to work with these files in any programming

language capable of working with HDF5 files would be easy. In contrast, the *.hic* format is not based on an existing standard, and using these files in a language not supported by the format authors would require good knowledge of binary storage. Tools to work with *.hic* files are written in Java, and while they benefit from high performance of this language, python, the language chosen by developers of *cooler* package and *cooltools*, is much more widely used in the scientific community. That's why I chose *.cool* as the format of choice to use in *coolpup.py*. Expanding my tool to support *.hic* could be one of the future directions of its development, however the existence of the *hic2cool* tool to convert *.hic* files to *.cool* files makes it not very necessary. A more interesting future direction could be supporting additional options for pileup analyses, such as inter-chromosomal pileups. For example, while CTCF/cohesin does not mediate interactions between chromosomes (Fudenberg et al., 2016; Rao et al., 2017b) it's very likely that cPRC1 binding sites can interact across different chromosomes, since interactions enrichment between them doesn't decrease with distance. Another feature already implemented in the latest version of *coolpup.py* but not described in the manuscript and Chapter, is calculation of pileups based on two sets of regions; this allows analysis of interactions between two different factors, or interactions between sites in different orientations, without generating huge files with all possible pairwise combinations of sites.

Deeper understanding of the biology functions of Polycomb bodies would require a combination of genetic manipulations and synthetic biology approaches. For example, deleting the CGI or preventing PRC1 binding at one

of the interacting partners could help reveal whether these interactions facilitate silencing. Alternatively, compaction and interactions could affect interactions of promoters with enhancers; this could be analysed by comparing gene activation through an enhancer after perturbing the other partner of an interaction. As a complementary approach, creating a strong cPRC1 binding site close to an endogenous target of cPRC1 should create a new interaction. It could then be determined whether this increases the robustness of silencing. Performing this using one of the key developmental regulators would facilitate demonstrating the functional relevance of Polycomb bodies – for example it would be possible to check whether differentiation of ES cells is affected by these manipulations.

Investigation into the potential existence of two distinct subclasses of PRC1 targets, detected in recent publications by Rob Klose's lab (Blackledge et al., 2019; Fursova et al., 2019) and correlated with regions that tend to have a robust enrichment for interactions with other RING1B targets, would be very important. While the majority of PRC1 targets absolutely rely on RING1B catalytic activity and vPRC1 for silencing, a small subset partially relies on cPRC1. Is there a difference in their establishment during mouse development *in vivo*? Are, perhaps, "loopers" found earlier, since they are on average much longer and have higher occupancy of PRC1 components (and in particular cPRC1)? It has long been known that expression of different PRC1 components is highly variable between different tissues in humans (Gunster et al., 2001): does this lead to different levels of PRC1-mediated interactions

between different cells types, and does it have any functional consequence? Are the “loopers” particularly affected by loss of cPRC1 components?

A very interesting further avenue of research would be the investigation of PRC1 binding dynamics during mouse early development, and PGC differentiation – during the waves of global DNA demethylation. The reorganization of PRC1 binding following changes in global levels of DNA methylation would presumably affect the 3D genome organization. The establishment of serum ES cell-like Polycomb binding pattern is interesting: does it occur later than the ICM stage, as would be predicted from very low level of DNA methylation and absence of detectable interactions in Hi-C data? Can it be established correctly in the absence of DNA methylation? Application of modern protein binding mapping technologies applicable to low cell numbers with high signal to noise, such as CUT&RUN or CUT&Tag (Kaya-Okur et al., 2019; Meers et al., 2019; Skene and Henikoff, 2017) would be required to investigate these *in vivo* stages, which would be interesting to combine with deeper Hi-C libraries than existing data (Du et al., 2017; Ke et al., 2017; Zhang et al., 2018). Comparison of the *in vivo* epigenome to that in ES cells derived in serum or 2i would help untangle the effects of “ground state” pluripotency, and those of adaptation to tissue culture conditions. For example, it might turn out that DNA methylation and Polycomb binding patterns in ES cells derived in 2i are more similar to those *in vivo* than in cells derived in serum, but grown in 2i.

Overall, the main questions for the field are linking the epigenome and the 3D genome with functionality. A lot of effort has been directed towards dissecting the role of TADs/CTCF-mediated loops in regulating enhancer-promoter

communication with often very modest impact (Paliou et al., 2019; Williamson et al., 2019). In other cases of more extensive genomic perturbations more robust effects of structure on gene expression could be observed (Despang et al., 2019; Franke et al., 2016; Lupiáñez et al., 2015). Forming general rules from these experiments is, however, impossible due to variability in results at different loci. Very little work has been performed concerning the functionality of other 3D structures, such as Polycomb-mediated compact domains or loops. Moreover, all work so far involved deletions of genomic sites involved in mediating interactions, while locally bringing or removing the effector proteins that mediate interactions could be more informative and would allow to avoid any side effects of genetic manipulations.

References

Abdennur, N., and Mirny, L.A. (2020). Cooler: scalable storage for Hi-C data and other genomically labeled arrays. Bioinformatics *36*, 311–316.

Abdennur, N., Schwarzer, W., Pekowska, A., Shaltiel, I.A., Huber, W., Haering, C.H., Mirny, L., and Spitz, F. (2018). Condensin II inactivation in interphase does not affect chromatin folding or gene expression. BioRxiv.

Abranches, E., Bekman, E., and Henrique, D. (2013). Generation and Characterization of a Novel Mouse Embryonic Stem Cell Line with a Dynamic Reporter of Nanog Expression. PLOS ONE *8*, e59928.

Alabert, C., Barth, T.K., Reverón-Gómez, N., Sidoli, S., Schmidt, A., Jensen, O.N., Imhof, A., and Groth, A. (2015). Two distinct modes for propagation of histone PTMs across the cell cycle. Genes Dev. *29*, 585–590.

Andrews, F.H., Strahl, B.D., and Kutateladze, T.G. (2016). Insights into newly discovered marks and readers of epigenetic information. Nat Chem Biol *12*, 662–668.

Antonysamy, S., Condon, B., Druzina, Z., Bonanno, J.B., Gheyi, T., Zhang, F., MacEwan, I., Zhang, A., Ashok, S., Rodgers, L., et al. (2013). Structural Context of Disease-Associated Mutations and Putative Mechanism of Autoinhibition Revealed by X-Ray Crystallographic Analysis of the EZH2-SET Domain. PLOS ONE *8*, e84147.

Atlasi, Y., Megchelenbrink, W., Peng, T., Habibi, E., Joshi, O., Wang, S.-Y., Wang, C., Logie, C., Poser, I., Marks, H., et al. (2019). Epigenetic modulation of a hardwired 3D chromatin landscape in two naive states of pluripotency. Nat Cell Biol *21*, 568–578.

Bantignies, F., Grimaud, C., Lavrov, S., Gabut, M., and Cavalli, G. (2003). Inheritance of Polycomb-dependent chromosomal interactions in Drosophila. Genes Dev *17*, 2406–2420.

Bantignies, F., Roure, V., Comet, I., Leblanc, B., Schuettengruber, B., Bonnet, J., Tixier, V., Mas, A., and Cavalli, G. (2011). Polycomb-dependent regulatory contacts between distant Hox loci in Drosophila. Cell *144*, 214–226.

Barau, J., Teissandier, A., Zamudio, N., Roy, S., Nalesso, V., Hérault, Y., Guillou, F., and Bourc'his, D. (2016). The DNA methyltransferase DNMT3C protects male germ cells from transposon activity. Science *354*, 909–912.

Barutcu, A.R., Fritz, A.J., Zaidi, S.K., van Wijnen, A.J., Lian, J.B., Stein, J.L., Nickerson, J.A., Imbalzano, A.N., and Stein, G.S. (2016). C-ing the Genome: A

Compendium of Chromosome Conformation Capture Methods to Study Higher-Order Chromatin Organization. J. Cell. Physiol. *231*, 31–35.

Beliveau, B.J., Joyce, E.F., Apostolopoulos, N., Yilmaz, F., Fonseka, C.Y., McCole, R.B., Chang, Y., Li, J.B., Senaratne, T.N., Williams, B.R., et al. (2012). Versatile design and synthesis platform for visualizing genomes with Oligopaint FISH probes. PNAS *109*, 21301–21306.

Bell, A.C., West, A.G., and Felsenfeld, G. (1999). The Protein CTCF Is Required for the Enhancer Blocking Activity of Vertebrate Insulators. Cell *98*, 387–396.

Bernstein, E., Duncan, E.M., Masui, O., Gil, J., Heard, E., and Allis, C.D. (2006). Mouse Polycomb Proteins Bind Differentially to Methylated Histone H3 and RNA and Are Enriched in Facultative Heterochromatin. Molecular and Cellular Biology *26*, 2560–2569.

Bestor, T.H., and Ingram, V.M. (1983). Two DNA methyltransferases from murine erythroleukemia cells: purification, sequence specificity, and mode of interaction with DNA. PNAS *80*, 5559–5563.

Bintu, B., Mateo, L.J., Su, J.-H., Sinnott-Armstrong, N.A., Parker, M., Kinrot, S., Yamaya, K., Boettiger, A.N., and Zhuang, X. (2018). Super-resolution chromatin tracing reveals domains and cooperative interactions in single cells. Science *362*, eaau1783.

Bird, A. (2002). DNA methylation patterns and epigenetic memory. Genes Dev. *16*, 6–21.

Bird, A. (2011). The Dinucleotide CG as a Genomic Signalling Module. Journal of Molecular Biology *409*, 47–53.

Bird, A., Taggart, M., Frommer, M., Miller, O.J., and Macleod, D. (1985). A fraction of the mouse genome that is derived from islands of nonmethylated, CpG-rich DNA. Cell *40*, 91–99.

Blackledge, N.P., Farcas, A.M., Kondo, T., King, H.W., McGouran, J.F., Hanssen, L.L.P., Ito, S., Cooper, S., Kondo, K., Koseki, Y., et al. (2014). Variant PRC1 Complex-Dependent H2A Ubiquitylation Drives PRC2 Recruitment and Polycomb Domain Formation. Cell *157*, 1445–1459.

Blackledge, N.P., Fursova, N.A., Kelley, J.R., Huseyin, M.K., Feldmann, A., and Klose, R.J. (2019). PRC1 catalytic activity is central to Polycomb system function. BioRxiv 667667.

Boettiger, A.N., Bintu, B., Moffitt, J.R., Wang, S., Beliveau, B.J., Fudenberg, G., Imakaev, M., Mirny, L.A., Wu, C., and Zhuang, X. (2016). Super-resolution

imaging reveals distinct chromatin folding for different epigenetic states. Nature *529*, 418–422.

Bonev, B., Mendelson Cohen, N., Szabo, Q., Fritsch, L., Papadopoulos, G.L., Lubling, Y., Xu, X., Lv, X., Hugnot, J.-P., Tanay, A., et al. (2017). Multiscale 3D Genome Rewiring during Mouse Neural Development. Cell *171*, 557-572.e24.

Bostick, M., Kim, J.K., Estève, P.-O., Clark, A., Pradhan, S., and Jacobsen, S.E. (2007). UHRF1 Plays a Role in Maintaining DNA Methylation in Mammalian Cells. Science *317*, 1760–1764.

Boyle, S., Gilchrist, S., Bridger, J.M., Mahy, N.L., Ellis, J.A., and Bickmore, W.A. (2001). The spatial organization of human chromosomes within the nuclei of normal and emerin-mutant cells. Hum Mol Genet *10*, 211–220.

Boyle, S., Rodesch, M.J., Halvensleben, H.A., Jeddeloh, J.A., and Bickmore, W.A. (2011). Fluorescence in situ hybridization with high-complexity repeat-free oligonucleotide probes generated by massively parallel synthesis. Chromosome Research *19*, 901–909.

Boyle, S., Flyamer, I.M., Williamson, I., Sengupta, D., Bickmore, W.A., and Illingworth, R.S. (2019). A Central Role for Canonical PRC1 in Shaping the 3D Nuclear Landscape. BioRxiv 2019.12.15.876771.

Bradley, A., Evans, M., Kaufman, M.H., and Robertson, E. (1984). Formation of germ-line chimaeras from embryo-derived teratocarcinoma cell lines. Nature *309*, 255–256.

Brinkman, A.B., Gu, H., Bartels, S.J.J., Zhang, Y., Matarese, F., Simmer, F., Marks, H., Bock, C., Gnirke, A., Meissner, A., et al. (2012). Sequential ChIP-bisulfite sequencing enables direct genome-scale investigation of chromatin and DNA methylation cross-talk. Genome Res. *22*, 1128–1138.

Buchenau, P., Hodgson, J., Strutt, H., and Arndt-Jovin, D.J. (1998). The Distribution of Polycomb-Group Proteins During Cell Division and Development in Drosophila Embryos: Impact on Models for Silencing. The Journal of Cell Biology *141*, 469–481.

Buchwald, G., van der Stoop, P., Weichenrieder, O., Perrakis, A., van Lohuizen, M., and Sixma, T.K. (2006). Structure and E3-ligase activity of the Ring–Ring complex of Polycomb proteins Bmi1 and Ring1b. EMBO J *25*, 2465–2474.

Burdon, T., Stracey, C., Chambers, I., Nichols, J., and Smith, A. (1999). Suppression of SHP-2 and ERK Signalling Promotes Self-Renewal of Mouse Embryonic Stem Cells. Developmental Biology *210*, 30–43.

Canela, A., Maman, Y., Jung, S., Wong, N., Callen, E., Day, A., Kieffer-Kwon, K.-R., Pekowska, A., Zhang, H., Rao, S.S.P., et al. (2017). Genome Organization Drives Chromosome Fragility. Cell *170*, 507-521.e18.

Canela, A., Maman, Y., Huang, S.N., Wutz, G., Tang, W., Zagnoli-Vieira, G., Callen, E., Wong, N., Day, A., Peters, J.-M., et al. (2019). Topoisomerase II-Induced Chromosome Breakage and Translocation Is Determined by Chromosome Architecture and Transcriptional Activity. Molecular Cell *75*, 252-266.e8.

Canham, M.A., Sharov, A.A., Ko, M.S.H., and Brickman, J.M. (2010). Functional Heterogeneity of Embryonic Stem Cells Revealed through Translational Amplification of an Early Endodermal Transcript. PLOS Biology *8*, e1000379.

Cao, R., and Zhang, Y. (2004). SUZ12 Is Required for Both the Histone Methyltransferase Activity and the Silencing Function of the EED-EZH2 Complex. Molecular Cell *15*, 57–67.

Cao, R., Wang, L., Wang, H., Xia, L., Erdjument-Bromage, H., Tempst, P., Jones, R.S., and Zhang, Y. (2002). Role of Histone H3 Lysine 27 Methylation in Polycomb-Group Silencing. Science *298*, 1039–1043.

Cardozo Gizzi, A.M., Cattoni, D.I., Fiche, J.-B., Espinola, S.M., Gurgo, J., Messina, O., Houbron, C., Ogiyama, Y., Papadopoulos, G.L., Cavalli, G., et al. (2019). Microscopy-Based Chromosome Conformation Capture Enables Simultaneous Visualization of Genome Organization and Transcription in Intact Organisms. Molecular Cell *74*, 212-222.e5.

Cartwright, P., McLean, C., Sheppard, A., Rivett, D., Jones, K., and Dalton, S. (2005). LIF/STAT3 controls ES cell self-renewal and pluripotency by a Myc-dependent mechanism. Development *132*, 885–896.

Casanova, M., Preissner, T., Cerase, A., Poot, R., Yamada, D., Li, X., Appanah, R., Bezstarosti, K., Demmers, J., Koseki, H., et al. (2011). Polycomblike 2 facilitates the recruitment of PRC2 Polycomb group complexes to the inactive X chromosome and to target loci in embryonic stem cells. Development *138*, 1471–1482.

Cassina, V., Manghi, M., Salerno, D., Tempestini, A., Iadarola, V., Nardo, L., Brioschi, S., and Mantegazza, F. (2016). Effects of cytosine methylation on DNA morphology: An atomic force microscopy study. Biochimica et Biophysica Acta (BBA) - General Subjects *1860*, 1–7.

Chédin, F., Lieber, M.R., and Hsieh, C.-L. (2002). The DNA methyltransferase-like protein DNMT3L stimulates de novo methylation by Dnmt3a. PNAS *99*, 16916–16921.

Chen, T., Ueda, Y., Dodge, J.E., Wang, Z., and Li, E. (2003). Establishment and Maintenance of Genomic Methylation Patterns in Mouse Embryonic Stem Cells by Dnmt3a and Dnmt3b. Molecular and Cellular Biology *23*, 5594–5605.

Chen, Y., Zhang, Y., Wang, Y., Zhang, L., Brinkman, E.K., Adam, S.A., Goldman, R., Steensel, B. van, Ma, J., and Belmont, A.S. (2018). Mapping 3D genome organization relative to nuclear compartments using TSA-Seq as a cytological ruler. J Cell Biol *217*, 4025–4048.

Chen, Z.-X., Mann, J.R., Hsieh, C.-L., Riggs, A.D., and Chédin, F. (2005). Physical and functional interactions between the human DNMT3L protein and members of the de novo methyltransferase family. Journal of Cellular Biochemistry *95*, 902–917.

Chuang, L.S.-H., Ian, H.-I., Koh, T.-W., Ng, H.-H., Xu, G., and Li, B.F.L. (1997). Human DNA-(Cytosine-5) Methyltransferase-PCNA Complex as a Target for p21WAF1. Science *277*, 1996–2000.

Cifuentes-Rojas, C., Hernandez, A.J., Sarma, K., and Lee, J.T. (2014). Regulatory Interactions between RNA and Polycomb Repressive Complex 2. Molecular Cell *55*, 171–185.

Cock, P.J.A., Antao, T., Chang, J.T., Chapman, B.A., Cox, C.J., Dalke, A., Friedberg, I., Hamelryck, T., Kauff, F., Wilczynski, B., et al. (2009). Biopython: freely available Python tools for computational molecular biology and bioinformatics. Bioinformatics *25*, 1422–1423.

Cohen, I., Zhao, D., Bar, C., Valdes, V.J., Dauber-Decker, K.L., Nguyen, M.B., Nakayama, M., Rendl, M., Bickmore, W.A., Koseki, H., et al. (2018). PRC1 Fine-tunes Gene Repression and Activation to Safeguard Skin Development and Stem Cell Specification. Cell Stem Cell *22*, 726-739.e7.

Coleman, R.T., and Struhl, G. (2017). Causal role for inheritance of H3K27me3 in maintaining the OFF state of a Drosophila HOX gene. Science *356*, eaai8236.

Cooper, D.N., Taggart, M.H., and Bird, A.P. (1983). Unmethlated domains in vertebrate DNA. Nucleic Acids Res *11*, 647–658.

Cooper, S., Dienstbier, M., Hassan, R., Schermelleh, L., Sharif, J., Blackledge, N.P., De Marco, V., Elderkin, S., Koseki, H., Klose, R., et al. (2014). Targeting Polycomb to Pericentric Heterochromatin in Embryonic Stem Cells Reveals a Role for H2AK119u1 in PRC2 Recruitment. Cell Reports *7*, 1456–1470.

Cooper, S., Grijzenhout, A., Underwood, E., Ancelin, K., Zhang, T., Nesterova, T.B., Anil-Kirmizitas, B., Bassett, A., Kooistra, S.M., Agger, K., et al. (2016).

Jarid2 binds mono-ubiquitylated H2A lysine 119 to mediate crosstalk between Polycomb complexes PRC1 and PRC2. Nature Communications *7*, 13661.

Crane, E., Bian, Q., McCord, R.P., Lajoie, B.R., Wheeler, B.S., Ralston, E.J., Uzawa, S., Dekker, J., and Meyer, B.J. (2015). Condensin-driven remodelling of X chromosome topology during dosage compensation. Nature *523*, 240–244.

Cremer, T., and Cremer, M. (2010). Chromosome Territories. Cold Spring Harb Perspect Biol *2*.

Cremer, M., von Hase, J., Volm, T., Brero, A., Kreth, G., Walter, J., Fischer, C., Solovei, I., Cremer, C., and Cremer, T. (2001). Non-random radial higher-order chromatin arrangements in nuclei of diploid human cells. Chromosome Res *9*, 541–567.

Cremer, M., Küpper, K., Wagler, B., Wizelman, L., Hase, J. v, Weiland, Y., Kreja, L., Diebold, J., Speicher, M.R., and Cremer, T. (2003). Inheritance of gene density–related higher order chromatin arrangements in normal and tumor cell nuclei. The Journal of Cell Biology *162*, 809–820.

Croft, J.A., Bridger, J.M., Boyle, S., Perry, P., Teague, P., and Bickmore, W.A. (1999). Differences in the Localization and Morphology of Chromosomes in the Human Nucleus. The Journal of Cell Biology *145*, 1119.

Cuadrado, A., Giménez-Llorente, D., Kojic, A., Rodríguez-Corsino, M., Cuartero, Y., Martín-Serrano, G., Gómez-López, G., Marti-Renom, M.A., and Losada, A. (2019). Specific Contributions of Cohesin-SA1 and Cohesin-SA2 to TADs and Polycomb Domains in Embryonic Stem Cells. Cell Reports *27*, 3500-3510.e4.

Czermin, B., Melfi, R., McCabe, D., Seitz, V., Imhof, A., and Pirrotta, V. (2002). Drosophila Enhancer of Zeste/ESC Complexes Have a Histone H3 Methyltransferase Activity that Marks Chromosomal Polycomb Sites. Cell *111*, 185–196.

Davidson, I.F., Bauer, B., Goetz, D., Tang, W., Wutz, G., and Peters, J.-M. (2019). DNA loop extrusion by human cohesin. Science *366*, 1338–1345.

Deaton, A.M., Gómez-Rodríguez, M., Mieczkowski, J., Tolstorukov, M.Y., Kundu, S., Sadreyev, R.I., Jansen, L.E., and Kingston, R.E. (2016). Enhancer regions show high histone H3.3 turnover that changes during differentiation. ELife *5*, e15316.

Dekker, J., Rippe, K., Dekker, M., and Kleckner, N. (2002). Capturing Chromosome Conformation. Science *295*, 1306–1311.

Denholtz, M., Bonora, G., Chronis, C., Splinter, E., de Laat, W., Ernst, J., Pellegrini, M., and Plath, K. (2013). Long-Range Chromatin Contacts in Embryonic Stem Cells Reveal a Role for Pluripotency Factors and Polycomb Proteins in Genome Organization. Cell Stem Cell *13*, 602–616.

Deniz, Ö., Frost, J.M., and Branco, M.R. (2019). Regulation of transposable elements by DNA modifications. Nat Rev Genet *20*, 417–431.

Despang, A., Schöpflin, R., Franke, M., Ali, S., Jerković, I., Paliou, C., Chan, W.-L., Timmermann, B., Wittler, L., Vingron, M., et al. (2019). Functional dissection of the Sox9 – Kcnj2 locus identifies nonessential and instructive roles of TAD architecture. Nat Genet *51*, 1263–1271.

Di Cerbo, V., Mohn, F., Ryan, D.P., Montellier, E., Kacem, S., Tropberger, P., Kallis, E., Holzner, M., Hoerner, L., Feldmann, A., et al. (2014). Acetylation of histone H3 at lysine 64 regulates nucleosome dynamics and facilitates transcription. ELife *3*, e01632.

Díaz, N., Kruse, K., Erdmann, T., Staiger, A.M., Ott, G., Lenz, G., and Vaquerizas, J.M. (2018). Chromatin conformation analysis of primary patient tissue using a low input Hi-C method. Nature Communications *9*, 4938.

Dixon, J.R., Selvaraj, S., Yue, F., Kim, A., Li, Y., Shen, Y., Hu, M., Liu, J.S., and Ren, B. (2012). Topological Domains in Mammalian Genomes Identified by Analysis of Chromatin Interactions. Nature *485*, 376–380.

Doble, B.W., Patel, S., Wood, G.A., Kockeritz, L.K., and Woodgett, J.R. (2007). Functional Redundancy of GSK-3α and GSK-3β in Wnt/β-Catenin Signaling Shown by Using an Allelic Series of Embryonic Stem Cell Lines. Developmental Cell *12*, 957–971.

Dostie, J., Richmond, T.A., Arnaout, R.A., Selzer, R.R., Lee, W.L., Honan, T.A., Rubio, E.D., Krumm, A., Lamb, J., Nusbaum, C., et al. (2006). Chromosome Conformation Capture Carbon Copy (5C): A massively parallel solution for mapping interactions between genomic elements. Genome Res *16*, 1299–1309.

Du, Q., Luu, P.-L., Stirzaker, C., and Clark, S.J. (2015). Methyl-CpG-binding domain proteins: readers of the epigenome. Epigenomics *7*, 1051–1073.

Du, Z., Zheng, H., Huang, B., Ma, R., Wu, J., Zhang, X., He, J., Xiang, Y., Wang, Q., Li, Y., et al. (2017). Allelic reprogramming of 3D chromatin architecture during early mammalian development. Nature *547*, 232–235.

Edwards, C.A., and Ferguson-Smith, A.C. (2007). Mechanisms regulating imprinted genes in clusters. Current Opinion in Cell Biology *19*, 281–289.

Edwards, J.R., Yarychkivska, O., Boulard, M., and Bestor, T.H. (2017). DNA methylation and DNA methyltransferases. Epigenetics & Chromatin *10*, 23.

Ehrlich, M., Gama-Sosa, M.A., Huang, L.-H., Midgett, R.M., Kuo, K.C., McCune, R.A., and Gehrke, C. (1982). Amount and distribution of 5-methylcytosine in human DNA from different types of tissues or cells. Nucleic Acids Res *10*, 2709–2721.

Elderkin, S., Maertens, G.N., Endoh, M., Mallery, D.L., Morrice, N., Koseki, H., Peters, G., Brockdorff, N., and Hiom, K. (2007). A Phosphorylated Form of Mel-18 Targets the Ring1B Histone H2A Ubiquitin Ligase to Chromatin. Molecular Cell *28*, 107–120.

Endoh, M., Endo, T.A., Endoh, T., Fujimura, Y., Ohara, O., Toyoda, T., Otte, A.P., Okano, M., Brockdorff, N., Vidal, M., et al. (2008). Polycomb group proteins Ring1A/B are functionally linked to the core transcriptional regulatory circuitry to maintain ES cell identity. Development *135*, 1513–1524.

Eskeland, R., Leeb, M., Grimes, G.R., Kress, C., Boyle, S., Sproul, D., Gilbert, N., Fan, Y., Skoultchi, A.I., Wutz, A., et al. (2010). Ring1B Compacts Chromatin Structure and Represses Gene Expression Independent of Histone Ubiquitination. Molecular Cell *38*, 452–464.

Evans, M.J., and Kaufman, M.H. (1981). Establishment in culture of pluripotential cells from mouse embryos. Nature *292*, 154–156.

Ewels, P.A., Peltzer, A., Fillinger, S., Alneberg, J., Patel, H., Wilm, A., Garcia, M.U., Tommaso, P.D., and Nahnsen, S. (2019). nf-core: Community curated bioinformatics pipelines. BioRxiv 610741.

Fabre, P.J., Benke, A., Joye, E., Nguyen Huynh, T.H., Manley, S., and Duboule, D. (2015). Nanoscale spatial organization of the HoxD gene cluster in distinct transcriptional states. Proc Natl Acad Sci U S A *112*, 13964–13969.

Falk, M., Feodorova, Y., Naumova, N., Imakaev, M., Lajoie, B.R., Leonhardt, H., Joffe, B., Dekker, J., Fudenberg, G., Solovei, I., et al. (2019). Heterochromatin drives compartmentalization of inverted and conventional nuclei. Nature *570*, 395–399.

Farcas, A.M., Blackledge, N.P., Sudbery, I., Long, H.K., McGouran, J.F., Rose, N.R., Lee, S., Sims, D., Cerase, A., Sheahan, T.W., et al. (2012). KDM2B links the Polycomb Repressive Complex 1 (PRC1) to recognition of CpG islands. ELife *1*, e00205.

Fatemi, M., Hermann, A., Gowher, H., and Jeltsch, A. (2002). Dnmt3a and Dnmt1 functionally cooperate during de novo methylation of DNA. European Journal of Biochemistry *269*, 4981–4984.

Fawcett, D.W. (1966). On the occurrence of a fibrous lamina on the inner aspect of the nuclear envelope in certain cells of vertebrates. American Journal of Anatomy *119*, 129–145.

Ferrari, K.J., Scelfo, A., Jammula, S., Cuomo, A., Barozzi, I., Stützer, A., Fischle, W., Bonaldi, T., and Pasini, D. (2014). Polycomb-Dependent H3K27me1 and H3K27me2 Regulate Active Transcription and Enhancer Fidelity. Molecular Cell *53*, 49–62.

Ficz, G., Hore, T.A., Santos, F., Lee, H.J., Dean, W., Arand, J., Krueger, F., Oxley, D., Paul, Y.-L., Walter, J., et al. (2013). FGF Signaling Inhibition in ESCs Drives Rapid Genome-wide Demethylation to the Epigenetic Ground State of Pluripotency. Cell Stem Cell *13*, 351–359.

Flyamer, I.M., Gassler, J., Imakaev, M., Brandão, H.B., Ulianov, S.V., Abdennur, N., Razin, S.V., Mirny, L.A., and Tachibana-Konwalski, K. (2017). Single-nucleus Hi-C reveals unique chromatin reorganization at oocyte-to-zygote transition. Nature *544*, 110–114.

Forcato, M., Nicoletti, C., Pal, K., Livi, C.M., Ferrari, F., and Bicciato, S. (2017). Comparison of computational methods for Hi-C data analysis. Nature Methods *14*, 679–685.

Franke, M., Ibrahim, D.M., Andrey, G., Schwarzer, W., Heinrich, V., Schöpflin, R., Kraft, K., Kempfer, R., Jerković, I., Chan, W.-L., et al. (2016). Formation of new chromatin domains determines pathogenicity of genomic duplications. Nature *538*, 265–269.

Frescas, D., Guardavaccaro, D., Bassermann, F., Koyama-Nasu, R., and Pagano, M. (2007). JHDM1B/FBXL10 is a nucleolar protein that represses transcription of ribosomal RNA genes. Nature *450*, 309–313.

Frey, F., Sheahan, T., Finkl, K., Stoehr, G., Mann, M., Benda, C., and Müller, J. (2016). Molecular basis of PRC1 targeting to Polycomb response elements by PhoRC. Genes Dev. *30*, 1116–1127.

Fudenberg, G., Imakaev, M., Lu, C., Goloborodko, A., Abdennur, N., and Mirny, L.A. (2016). Formation of Chromosomal Domains by Loop Extrusion. Cell Reports *15*, 2038–2049.

Fursova, N.A., Blackledge, N.P., Nakayama, M., Ito, S., Koseki, Y., Farcas, A.M., King, H.W., Koseki, H., and Klose, R.J. (2019). Synergy between Variant PRC1 Complexes Defines Polycomb-Mediated Gene Repression. Molecular Cell *0*.

Fussner, E., Strauss, M., Djuric, U., Li, R., Ahmed, K., Hart, M., Ellis, J., and Bazett-Jones, D.P. (2012). Open and closed domains in the mouse genome are configured as 10-nm chromatin fibres. EMBO Rep *13*, 992–996.

Galganski, L., Urbanek, M.O., and Krzyzosiak, W.J. (2017). Nuclear speckles: molecular organization, biological function and role in disease. Nucleic Acids Res *45*, 10350–10368.

Gao, Z., Zhang, J., Bonasio, R., Strino, F., Sawai, A., Parisi, F., Kluger, Y., and Reinberg, D. (2012). PCGF homologs, CBX proteins, and RYBP define functionally distinct PRC1 family complexes. Mol. Cell *45*, 344–356.

Gardiner-Garden, M., and Frommer, M. (1987). CpG Islands in vertebrate genomes. Journal of Molecular Biology *196*, 261–282.

Gassler, J., Brandão, H.B., Imakaev, M., Flyamer, I.M., Ladstätter, S., Bickmore, W.A., Peters, J., Mirny, L.A., and Tachibana, K. (2017). A mechanism of cohesin-dependent loop extrusion organizes zygotic genome architecture. EMBO J *36*, 3600–3618.

Gavrilov, A.A., Gushchanskaya, E.S., Strelkova, O., Zhironkina, O., Kireev, I.I., Iarovaia, O.V., and Razin, S.V. (2013). Disclosure of a structural milieu for the proximity ligation reveals the elusive nature of an active chromatin hub. Nucleic Acids Res *41*, 3563–3575.

Ghirlando, R., and Felsenfeld, G. (2016). CTCF: making the right connections. Genes Dev. *30*, 881–891.

Gothe, H.J., Bouwman, B.A.M., Gusmao, E.G., Piccinno, R., Petrosino, G., Sayols, S., Drechsel, O., Minneker, V., Josipovic, N., Mizi, A., et al. (2019). Spatial Chromosome Folding and Active Transcription Drive DNA Fragility and Formation of Oncogenic MLL Translocations. Molecular Cell *75*, 267-283.e12.

Gowher, H., Liebert, K., Hermann, A., Xu, G., and Jeltsch, A. (2005). Mechanism of Stimulation of Catalytic Activity of Dnmt3A and Dnmt3B DNA-(cytosine-C5)-methyltransferases by Dnmt3L. J. Biol. Chem. *280*, 13341–13348.

Grabole, N., Tischler, J., Hackett, J.A., Kim, S., Tang, F., Leitch, H.G., Magnúsdóttir, E., and Surani, M.A. (2013). Prdm14 promotes germline fate and naive pluripotency by repressing FGF signalling and DNA methylation. EMBO Reports *14*, 629–637.

Grau, D.J., Chapman, B.A., Garlick, J.D., Borowsky, M., Francis, N.J., and Kingston, R.E. (2011). Compaction of chromatin by diverse Polycomb group proteins requires localized regions of high charge. Genes Dev. *25*, 2210–2221.

Gruenbaum, Y., Cedar, H., and Razin, A. (1982). Substrate and sequence specificity of a eukaryotic DNA methylase. Nature *295*, 620–622.

Gunster, M.J., Raaphorst, F.M., Hamer, K.M., Blaauwen, J.L. den, Fieret, E., Meijer, C.J.L.M., and Otte, A.P. (2001). Differential expression of human Polycomb group proteins in various tissues and cell types. Journal of Cellular Biochemistry *81*, 129–143.

Guo, F., Li, X., Liang, D., Li, T., Zhu, P., Guo, H., Wu, X., Wen, L., Gu, T.-P., Hu, B., et al. (2014). Active and Passive Demethylation of Male and Female Pronuclear DNA in the Mammalian Zygote. Cell Stem Cell *15*, 447–459.

Habibi, E., Brinkman, A.B., Arand, J., Kroeze, L.I., Kerstens, H.H.D., Matarese, F., Lepikhov, K., Gut, M., Brun-Heath, I., Hubner, N.C., et al. (2013). Whole-Genome Bisulfite Sequencing of Two Distinct Interconvertible DNA Methylomes of Mouse Embryonic Stem Cells. Cell Stem Cell *13*, 360–369.

Hackett, J.A., Reddington, J.P., Nestor, C.E., Dunican, D.S., Branco, M.R., Reichmann, J., Reik, W., Surani, M.A., Adams, I.R., and Meehan, R.R. (2012). Promoter DNA methylation couples genome-defence mechanisms to epigenetic reprogramming in the mouse germline. Development *139*, 3623–3632.

Hackett, J.A., Dietmann, S., Murakami, K., Down, T.A., Leitch, H.G., and Surani, M.A. (2013). Synergistic Mechanisms of DNA Demethylation during Transition to Ground-State Pluripotency. Stem Cell Reports *1*, 518–531.

Hall, J., Guo, G., Wray, J., Eyres, I., Nichols, J., Grotewold, L., Morfopoulou, S., Humphreys, P., Mansfield, W., Walker, R., et al. (2009). Oct4 and LIF/Stat3 Additively Induce Krüppel Factors to Sustain Embryonic Stem Cell Self-Renewal. Cell Stem Cell *5*, 597–609.

Hata, K., Okano, M., Lei, H., and Li, E. (2002). Dnmt3L cooperates with the Dnmt3 family of de novo DNA methyltransferases to establish maternal imprints in mice. Development *129*, 1983–1993.

Hayashi, K., Lopes, S.M.C. de S., Tang, F., and Surani, M.A. (2008). Dynamic Equilibrium and Heterogeneity of Mouse Pluripotent Stem Cells with Distinct Functional and Epigenetic States. Cell Stem Cell *3*, 391–401.

He, J., Shen, L., Wan, M., Taranova, O., Wu, H., and Zhang, Y. (2013). Kdm2b maintains murine embryonic stem cell status by recruiting PRC1 complex to CpG islands of developmental genes. Nature Cell Biology *15*, 373–384.

Hendrich, B., Hardeland, U., Ng, H.-H., Jiricny, J., and Bird, A. (1999). The thymine glycosylase MBD4 can bind to the product of deamination at methylated CpG sites. Nature *401*, 301–304.

Henikoff, S., and Smith, M.M. (2015). Histone Variants and Epigenetics. Cold Spring Harb Perspect Biol *7*, a019364.

Højfeldt, J.W., Laugesen, A., Willumsen, B.M., Damhofer, H., Hedehus, L., Tvardovskiy, A., Mohammad, F., Jensen, O.N., and Helin, K. (2018). Accurate H3K27 methylation can be established de novo by SUZ12-directed PRC2. Nat Struct Mol Biol *25*, 225–232.

Holliday, R., and Pugh, J.E. (1975). DNA modification mechanisms and gene activity during development. Science *187*, 226–232.

Hsieh, T.-H.S., Weiner, A., Lajoie, B., Dekker, J., Friedman, N., and Rando, O.J. (2015). Mapping Nucleosome Resolution Chromosome Folding in Yeast by Micro-C. Cell *162*, 108–119.

Hsieh, T.-H.S., Fudenberg, G., Goloborodko, A., and Rando, O.J. (2016). Micro-C XL: assaying chromosome conformation from the nucleosome to the entire genome. Nat Meth *13*, 1009–1011.

Hsieh, T.-H.S., Slobodyanyuk, E., Hansen, A.S., Cattoglio, C., Rando, O.J., Tjian, R., and Darzacq, X. (2019). Resolving the 3D landscape of transcription-linked mammalian chromatin folding. BioRxiv 638775.

Hunter, J.D. (2007). Matplotlib: A 2D Graphics Environment. Computing in Science Engineering *9*, 90–95.

Hurlin, P.J. (1999). Mga, a dual-specificity transcription factor that interacts with Max and contains a T-domain DNA-binding motif. The EMBO Journal *18*, 7019–7028.

ter Huurne, M., Chappell, J., Dalton, S., and Stunnenberg, H.G. (2017). Distinct Cell-Cycle Control in Two Different States of Mouse Pluripotency. Cell Stem Cell *21*, 449-455.e4.

Iida, T., Suetake, I., Tajima, S., Morioka, H., Ohta, S., Obuse, C., and Tsurimoto, T. (2002). PCNA clamp facilitates action of DNA cytosine methyltransferase 1 on hemimethylated DNA. Genes to Cells *7*, 997–1007.

Illingworth, R.S. (2019). Chromatin folding and nuclear architecture: PRC1 function in 3D. Current Opinion in Genetics & Development *55*, 82–90.

Illingworth, R.S., and Bird, A.P. (2009). CpG islands – 'A rough guide.' FEBS Letters *583*, 1713–1720.

Illingworth, R.S., Gruenewald-Schneider, U., Webb, S., Kerr, A.R.W., James, K.D., Turner, D.J., Smith, C., Harrison, D.J., Andrews, R., and Bird, A.P.

(2010). Orphan CpG Islands Identify Numerous Conserved Promoters in the Mammalian Genome. PLOS Genetics *6*, e1001134.

Illingworth, R.S., Botting, C.H., Grimes, G.R., Bickmore, W.A., and Eskeland, R. (2012). PRC1 and PRC2 Are Not Required for Targeting of H2A.Z to Developmental Genes in Embryonic Stem Cells. PLOS ONE *7*, e34848.

Illingworth, R.S., Moffat, M., Mann, A.R., Read, D., Hunter, C.J., Pradeepa, M.M., Adams, I.R., and Bickmore, W.A. (2015). The E3 ubiquitin ligase activity of RING1B is not essential for early mouse development. Genes Dev *29*, 1897–1902.

Imakaev, M., Fudenberg, G., McCord, R.P., Naumova, N., Goloborodko, A., Lajoie, B.R., Dekker, J., and Mirny, L.A. (2012). Iterative correction of Hi-C data reveals hallmarks of chromosome organization. Nature Methods *9*, 999–1003.

Isono, K., Fujimura, Y., Shinga, J., Yamaki, M., O-Wang, J., Takihara, Y., Murahashi, Y., Takada, Y., Mizutani-Koseki, Y., and Koseki, H. (2005). Mammalian Polyhomeotic Homologues Phc2 and Phc1 Act in Synergy To Mediate Polycomb Repression of Hox Genes. Mol Cell Biol *25*, 6694–6706.

Isono, K., Endo, T.A., Ku, M., Yamada, D., Suzuki, R., Sharif, J., Ishikura, T., Toyoda, T., Bernstein, B.E., and Koseki, H. (2013). SAM Domain Polymerization Links Subnuclear Clustering of PRC1 to Gene Silencing. Developmental Cell *26*, 565–577.

Janzer, A., Stamm, K., Becker, A., Zimmer, A., Buettner, R., and Kirfel, J. (2012). The H3K4me3 Histone Demethylase Fbxl10 Is a Regulator of Chemokine Expression, Cellular Morphology, and the Metabolome of Fibroblasts. J. Biol. Chem. *287*, 30984–30992.

Jermann, P., Hoerner, L., Burger, L., and Schübeler, D. (2014). Short sequences can efficiently recruit histone H3 lysine 27 trimethylation in the absence of enhancer activity and DNA methylation. PNAS *111*, E3415–E3421.

Jiao, L., and Liu, X. (2015). Structural basis of histone H3K27 trimethylation by an active polycomb repressive complex 2. Science *350*, aac4383.

Jones, P.A. (2012). Functions of DNA methylation: islands, start sites, gene bodies and beyond. Nature Reviews Genetics *13*, 484–492.

Jones, P.A., and Liang, G. (2009). Rethinking how DNA methylation patterns are maintained. Nature Reviews Genetics *10*, 805–811.

Jones, E., Oliphant, T., Peterson, P., and others (2001). SciPy: Open source scientific tools for Python.

Joshi, O., Wang, S.-Y., Kuznetsova, T., Atlasi, Y., Peng, T., Fabre, P.J., Habibi, E., Shaik, J., Saeed, S., Handoko, L., et al. (2015). Dynamic Reorganization of Extremely Long-Range Promoter-Promoter Interactions between Two States of Pluripotency. Cell Stem Cell *17*, 748–757.

Jung, H.R., Pasini, D., Helin, K., and Jensen, O.N. (2010). Quantitative Mass Spectrometry of Histones H3.2 and H3.3 in Suz12-deficient Mouse Embryonic Stem Cells Reveals Distinct, Dynamic Post-translational Modifications at Lys-27 and Lys-36. Molecular & Cellular Proteomics *9*, 838–850.

Jung, Y.H., Sauria, M.E.G., Lyu, X., Cheema, M.S., Ausio, J., Taylor, J., and Corces, V.G. (2017). Chromatin States in Mouse Sperm Correlate with Embryonic and Adult Regulatory Landscapes. Cell Rep *18*, 1366–1382.

Kagiwada, S., Kurimoto, K., Hirota, T., Yamaji, M., and Saitou, M. (2013). Replication-coupled passive DNA demethylation for the erasure of genome imprints in mice. The EMBO Journal *32*, 340–353.

Kalb, R., Latwiel, S., Baymaz, H.I., Jansen, P.W.T.C., Müller, C.W., Vermeulen, M., and Müller, J. (2014). Histone H2A monoubiquitination promotes histone H3 methylation in Polycomb repression. Nature Structural & Molecular Biology *21*, 569–571.

Kaneko, S., Son, J., Bonasio, R., Shen, S.S., and Reinberg, D. (2014). Nascent RNA interaction keeps PRC2 activity poised and in check. Genes Dev. *28*, 1983–1988.

Kang, H., McElroy, K.A., Jung, Y.L., Alekseyenko, A.A., Zee, B.M., Park, P.J., and Kuroda, M.I. (2015). Sex comb on midleg (Scm) is a functional link between PcG-repressive complexes in Drosophila. Genes Dev. *29*, 1136–1150.

Kaustov, L., Ouyang, H., Amaya, M., Lemak, A., Nady, N., Duan, S., Wasney, G.A., Li, Z., Vedadi, M., Schapira, M., et al. (2011). Recognition and Specificity Determinants of the Human Cbx Chromodomains. J. Biol. Chem. *286*, 521–529.

Kawasaki, Y., Lee, J., Matsuzawa, A., Kohda, T., Kaneko-Ishino, T., and Ishino, F. (2014). Active DNA demethylation is required for complete imprint erasure in primordial germ cells. Scientific Reports *4*, 3658.

Kaya-Okur, H.S., Wu, S.J., Codomo, C.A., Pledger, E.S., Bryson, T.D., Henikoff, J.G., Ahmad, K., and Henikoff, S. (2019). CUT&Tag for efficient epigenomic profiling of small samples and single cells. Nat Commun *10*, 1–10.

Ke, Y., Xu, Y., Chen, X., Feng, S., Liu, Z., Sun, Y., Yao, X., Li, F., Zhu, W., Gao, L., et al. (2017). 3D Chromatin Structures of Mature Gametes and

Structural Reprogramming during Mammalian Embryogenesis. Cell *170*, 367-381.e20.

Kim, C.A., and Bowie, J.U. (2003). SAM domains: uniform structure, diversity of function. Trends in Biochemical Sciences *28*, 625–628.

Kim, C.A., Gingery, M., Pilpa, R.M., and Bowie, J.U. (2002). The SAM domain of polyhomeotic forms a helical polymer. Nat Struct Mol Biol *9*, 453–457.

Kim, U.J., Shizuya, H., de Jong, P.J., Birren, B., and Simon, M.I. (1992). Stable propagation of cosmid sized human DNA inserts in an F factor based vector. Nucleic Acids Res. *20*, 1083–1085.

Klutstein, M., Nejman, D., Greenfield, R., and Cedar, H. (2016). DNA Methylation in Cancer and Aging. Cancer Res *76*, 3446–3450.

Koopman, P., and Cotton, R.G.H. (1984). A factor produced by feeder cells which inhibits embryonal carcinoma cell differentiation: Characterization and partial purification. Experimental Cell Research *154*, 233–242.

Korthauer, K., and Irizarry, R.A. (2018). Genome-wide repressive capacity of promoter DNA methylation is revealed through epigenomic manipulation. BioRxiv 381145.

Koyama-Nasu, R., David, G., and Tanese, N. (2007). The F-box protein Fbl10 is a novel transcriptional repressor of c-Jun. Nature Cell Biology *9*, 1074–1080.

Krietenstein, N., Abraham, S., Venev, S.V., Abdennur, N., Gibcus, J., Hsieh, T.-H.S., Parsi, K.M., Yang, L., Maehr, R., Mirny, L.A., et al. (2019). Ultrastructural details of mammalian chromosome architecture. BioRxiv 639922.

Kruse, K., Diaz, N., Enriquez-Gasca, R., Gaume, X., Torres-Padilla, M.-E., and Vaquerizas, J.M. (2019). Transposable elements drive reorganisation of 3D chromatin during early embryogenesis. BioRxiv 523712.

Ku, M., Koche, R.P., Rheinbay, E., Mendenhall, E.M., Endoh, M., Mikkelsen, T.S., Presser, A., Nusbaum, C., Xie, X., Chi, A.S., et al. (2008). Genomewide analysis of PRC1 and PRC2 occupancy identifies two classes of bivalent domains. PLoS Genet. *4*, e1000242.

Kunath, T., Saba-El-Leil, M.K., Almousailleakh, M., Wray, J., Meloche, S., and Smith, A. (2007). FGF stimulation of the Erk1/2 signalling cascade triggers transition of pluripotent embryonic stem cells from self-renewal to lineage commitment. Development *134*, 2895–2902.

Kundu, S., Ji, F., Sunwoo, H., Jain, G., Lee, J.T., Sadreyev, R.I., Dekker, J., and Kingston, R.E. (2017). Polycomb Repressive Complex 1 Generates

Discrete Compacted Domains that Change during Differentiation. Molecular Cell *65*, 432-446.e5.

Lajoie, B.R., Dekker, J., and Kaplan, N. (2015). The Hitchhiker's guide to Hi-C analysis: Practical guidelines. Methods *72*, 65–75.

Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., et al. (2001). Initial sequencing and analysis of the human genome. Nature *409*, 860–921.

Laprell, F., Finkl, K., and Müller, J. (2017). Propagation of Polycomb-repressed chromatin requires sequence-specific recruitment to DNA. Science *356*, 85–88.

Larsen, F., Gundersen, G., Lopez, R., and Prydz, H. (1992). CpG islands as gene markers in the human genome. Genomics *13*, 1095–1107.

Larson, A.G., Elnatan, D., Keenen, M.M., Trnka, M.J., Johnston, J.B., Burlingame, A.L., Agard, D.A., Redding, S., and Narlikar, G.J. (2017). Liquid droplet formation by HP1α suggests a role for phase separation in heterochromatin. Nature *547*, 236–240.

Lau, M.S., Schwartz, M.G., Kundu, S., Savol, A.J., Wang, P.I., Marr, S.K., Grau, D.J., Schorderet, P., Sadreyev, R.I., Tabin, C.J., et al. (2017). Mutation of a nucleosome compaction region disrupts Polycomb-mediated axial patterning. Science *355*, 1081–1084.

Lea, A.J., Vockley, C.M., Johnston, R.A., Del Carpio, C.A., Barreiro, L.B., Reddy, T.E., and Tung, J. (2018). Genome-wide quantification of the effects of DNA methylation on human gene regulation. ELife *7*, e37513.

Lee, C.-H., Holder, M., Grau, D., Saldaña-Meyer, R., Yu, J.-R., Ganai, R.A., Zhang, J., Wang, M., LeRoy, G., Dobenecker, M.-W., et al. (2018). Distinct Stimulatory Mechanisms Regulate the Catalytic Activity of Polycomb Repressive Complex 2. Molecular Cell *70*, 435-448.e5.

Lee, H.J., Hore, T.A., and Reik, W. (2014). Reprogramming the Methylome: Erasing Memory and Creating Diversity. Cell Stem Cell *14*, 710–719.

Leeb, M., and Wutz, A. (2007). Ring1B is crucial for the regulation of developmental control genes and PRC1 proteins but not X inactivation in embryonic cells. The Journal of Cell Biology *178*, 219–229.

Leitch, H.G., McEwen, K.R., Turp, A., Encheva, V., Carroll, T., Grabole, N., Mansfield, W., Nashun, B., Knezovich, J.G., Smith, A., et al. (2013). Naive pluripotency is associated with global DNA hypomethylation. Nat. Struct. Mol. Biol. *20*, 311–316.

Lekschas, F., Bach, B., Kerpedjiev, P., Gehlenborg, N., and Pfister, H. (2018). HiPiler: Visual Exploration of Large Genome Interaction Matrices with Interactive Small Multiples. IEEE Transactions on Visualization and Computer Graphics *24*, 522–531.

Levine, S.S., Weiss, A., Erdjument-Bromage, H., Shao, Z., Tempst, P., and Kingston, R.E. (2002). The Core of the Polycomb Repressive Complex Is Compositionally and Functionally Conserved in Flies and Humans. Molecular and Cellular Biology *22*, 6070–6078.

Lewis, E.B. (1978). A gene complex controlling segmentation in Drosophila. Nature *276*, 565–570.

Lewis, P.H. (1947). Drosophila Information Service *21*, 69.

Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. ArXiv:1303.3997 [q-Bio].

Li, G., Margueron, R., Ku, M., Chambon, P., Bernstein, B.E., and Reinberg, D. (2010). Jarid2 and PRC2, partners in regulating gene expression. Genes Dev. *24*, 368–380.

Li, H., Liefke, R., Jiang, J., Kurland, J.V., Tian, W., Deng, P., Zhang, W., He, Q., Patel, D.J., Bulyk, M.L., et al. (2017). Polycomb-like proteins link the PRC2 complex to CpG islands. Nature *549*, 287–291.

Li, Y., Haarhuis, J.H.I., Cacciatore, Á.S., Oldenkamp, R., Ruiten, M.S. van, Willems, L., Teunissen, H., Muir, K.W., Wit, E. de, Rowland, B.D., et al. (2020). The structural basis for cohesin–CTCF-anchored loops. Nature 1–9.

Liang, G., Chan, M.F., Tomigahara, Y., Tsai, Y.C., Gonzales, F.A., Li, E., Laird, P.W., and Jones, P.A. (2002). Cooperativity between DNA Methyltransferases in the Maintenance Methylation of Repetitive Elements. Molecular and Cellular Biology *22*, 480–491.

Lieberman-Aiden, E., van Berkum, N.L., Williams, L., Imakaev, M., Ragoczy, T., Telling, A., Amit, I., Lajoie, B.R., Sabo, P.J., Dorschner, M.O., et al. (2009). Comprehensive mapping of long range interactions reveals folding principles of the human genome. Science *326*, 289–293.

Liu, X., Wang, C., Liu, W., Li, J., Li, C., Kou, X., Chen, J., Zhao, Y., Gao, H., Wang, H., et al. (2016). Distinct features of H3K4me3 and H3K27me3 chromatin domains in pre-implantation embryos. Nature *537*, 558–562.

Long, H.K., King, H.W., Patient, R.K., Odom, D.T., and Klose, R.J. (2016). Protection of CpG islands from DNA methylation is DNA-encoded and evolutionarily conserved. Nucleic Acids Res *44*, 6693–6706.

Luger, K., Mäder, A.W., Richmond, R.K., Sargent, D.F., and Richmond, T.J. (1997). Crystal structure of the nucleosome core particle at 2.8 A resolution. Nature *389*, 18.

Lupiáñez, D.G., Kraft, K., Heinrich, V., Krawitz, P., Brancati, F., Klopocki, E., Horn, D., Kayserili, H., Opitz, J.M., Laxova, R., et al. (2015). Disruptions of Topological Chromatin Domains Cause Pathogenic Rewiring of Gene-Enhancer Interactions. Cell *161*, 1012–1025.

Lyko, F. (2018). The DNA methyltransferase family: a versatile toolkit for epigenetic regulation. Nature Reviews Genetics *19*, 81–92.

Ma, W., Ay, F., Lee, C., Gulsoy, G., Deng, X., Cook, S., Hesson, J., Cavanaugh, C., Ware, C.B., Krumm, A., et al. (2015). Fine-scale chromatin interaction maps reveal the cis-regulatory landscape of human lincRNA genes. Nature Methods *12*, 71–78.

Ma, Z., Swigut, T., Valouev, A., Rada-Iglesias, A., and Wysocka, J. (2011). Sequence-specific regulator Prdm14 safeguards mouse ESCs from entering extraembryonic endoderm fates. Nature Structural & Molecular Biology *18*, 120–127.

Malaguti, M., Nistor, P.A., Blin, G., Pegg, A., Zhou, X., and Lowell, S. (2013). Bone morphogenic protein signalling suppresses differentiation of pluripotent cells by maintaining expression of E-Cadherin. ELife *2*, e01197.

Margueron, R., Li, G., Sarma, K., Blais, A., Zavadil, J., Woodcock, C.L., Dynlacht, B.D., and Reinberg, D. (2008). Ezh1 and Ezh2 Maintain Repressive Chromatin through Different Mechanisms. Molecular Cell *32*, 503–518.

Margueron, R., Justin, N., Ohno, K., Sharpe, M.L., Son, J., Drury Iii, W.J., Voigt, P., Martin, S.R., Taylor, W.R., De Marco, V., et al. (2009). Role of the polycomb protein EED in the propagation of repressive histone marks. Nature *461*, 762–767.

Marks, H., Kalkan, T., Menafra, R., Denissov, S., Jones, K., Hofemeister, H., Nichols, J., Kranz, A., Francis Stewart, A., Smith, A., et al. (2012). The Transcriptional and Epigenomic Foundations of Ground State Pluripotency. Cell *149*, 590–604.

Martin, G.R. (1981). Isolation of a pluripotent cell line from early mouse embryos cultured in medium conditioned by teratocarcinoma stem cells. PNAS *78*, 7634–7638.

Mateo, L.J., Murphy, S.E., Hafner, A., Cinquini, I.S., Walker, C.A., and Boettiger, A.N. (2019). Visualizing DNA folding and RNA in embryos at single-cell resolution. Nature *568*, 49–54.

Matheson, L., and Elderkin, S. (2018). 13 - Polycomb Bodies. In Nuclear Architecture and Dynamics, C. Lavelle, and J.-M. Victor, eds. (Boston: Academic Press), pp. 297–320.

McKinney, W. (2010). Data Structures for Statistical Computing in Python. 6.

McLaughlin, K.A. (2018). Role of DNA methylation and Polycomb machineries in directing higher-order chromatin architecture in embryonic stem cell. University of Edinburgh.

McLaughlin, K.A., Flyamer, I.M., Thomson, J.P., Mjoseng, H.K., Shukla, R., Williamson, I., Grimes, G.R., Illingworth, R.S., Adams, I.R., Pennings, S., et al. (2019). DNA methylation directs polycomb-dependent 3D genome re-organisation in naive pluripotency. BioRxiv 527309.

Meehan, R.R., Lewis, J.D., McKay, S., Kleiner, E.L., and Bird, A.P. (1989). Identification of a mammalian protein that binds specifically to DNA containing methylated CpGs. Cell *58*, 499–507.

Meers, M.P., Bryson, T.D., Henikoff, J.G., and Henikoff, S. (2019). Improved CUT&RUN chromatin profiling tools. ELife *8*, e46314.

Mendenhall, E.M., Koche, R.P., Truong, T., Zhou, V.W., Issac, B., Chi, A.S., Ku, M., and Bernstein, B.E. (2010). GC-Rich Sequence Elements Recruit PRC2 in Mammalian ES Cells. PLOS Genetics *6*, e1001244.

Michael Waskom, Olga Botvinnik, Drew O'Kane, Paul Hobson, Joel Ostblom, Saulius Lukauskas, David C Gemperline, Tom Augspurger, Yaroslav Halchenko, John B. Cole, et al. (2018). mwaskom/seaborn: v0.9.0 (July 2018) (Zenodo).

Mirny, L.A., Imakaev, M., and Abdennur, N. (2019). Two major mechanisms of chromosome organization. Current Opinion in Cell Biology *58*, 142–152.

Morey, L., Santanach, A., Blanco, E., Aloia, L., Nora, E.P., Bruneau, B.G., and Di Croce, L. (2015). Polycomb Regulates Mesoderm Cell Fate-Specification in Embryonic Stem Cells through Activation and Repression Mechanisms. Cell Stem Cell *17*, 300–315.

Müller, J., Hart, C.M., Francis, N.J., Vargas, M.L., Sengupta, A., Wild, B., Miller, E.L., O'Connor, M.B., Kingston, R.E., and Simon, J.A. (2002). Histone Methyltransferase Activity of a Drosophila Polycomb Group Repressor Complex. Cell *111*, 197–208.

Nagano, T., Lubling, Y., Stevens, T.J., Schoenfelder, S., Yaffe, E., Dean, W., Laue, E.D., Tanay, A., and Fraser, P. (2013). Single-cell Hi-C reveals cell-to-cell variability in chromosome structure. Nature *502*, 59–64.

Nagano, T., Lubling, Y., Várnai, C., Dudley, C., Leung, W., Baran, Y., Mendelson Cohen, N., Wingett, S., Fraser, P., and Tanay, A. (2017). Cell-cycle dynamics of chromosomal organization at single-cell resolution. Nature *547*, 61–67.

Naumova, N., Imakaev, M., Fudenberg, G., Zhan, Y., Lajoie, B.R., Mirny, L.A., and Dekker, J. (2013). Organization of the Mitotic Chromosome. Science *342*, 948–953.

Neira, J.L., Román-Trufero, M., Contreras, L.M., Prieto, J., Singh, G., Barrera, F.N., Renart, M.L., and Vidal, M. (2009). The Transcriptional Repressor RYBP Is a Natively Unfolded Protein Which Folds upon Binding to DNA. Biochemistry *48*, 1348–1360.

Nekrasov, M., Wild, B., and Müller, J. (2005). Nucleosome binding and histone methyltransferase activity of Drosophila PRC2. EMBO Reports *6*, 348–353.

Nir, G., Farabella, I., Estrada, C.P., Ebeling, C.G., Beliveau, B.J., Sasaki, H.M., Lee, S.D., Nguyen, S.C., McCole, R.B., Chattoraj, S., et al. (2018). Walking along chromosomes with super-resolution imaging, contact maps, and integrative modeling. PLOS Genetics *14*, e1007872.

Niwa, H., Ogawa, K., Shimosato, D., and Adachi, K. (2009). A parallel circuit of LIF signalling pathways maintains pluripotency of mouse ES cells. Nature *460*, 118–122.

Nora, E.P., Lajoie, B.R., Schulz, E.G., Giorgetti, L., Okamoto, I., Servant, N., Piolot, T., van Berkum, N.L., Meisig, J., Sedat, J., et al. (2012). Spatial partitioning of the regulatory landscape of the X-inactivation centre. Nature *485*, 381–385.

Nora, E.P., Goloborodko, A., Valton, A.-L., Gibcus, J.H., Uebersohn, A., Abdennur, N., Dekker, J., Mirny, L.A., and Bruneau, B.G. (2017). Targeted Degradation of CTCF Decouples Local Insulation of Chromosome Domains from Genomic Compartmentalization. Cell *169*, 930-944.e22.

Nora, E.P., Caccianini, L., Fudenberg, G., Kameswaran, V., Nagle, A., Uebersohn, A., So, K., Hajj, B., Saux, A.L., Coulon, A., et al. (2019). Molecular basis of CTCF binding polarity in genome folding. BioRxiv 2019.12.13.876177.

Ohno, R., Nakayama, M., Naruse, C., Okashita, N., Takano, O., Tachibana, M., Asano, M., Saitou, M., and Seki, Y. (2013). A replication-dependent passive mechanism modulates DNA demethylation in mouse primordial germ cells. Development *140*, 2892–2903.

Okano, M., Bell, D.W., Haber, D.A., and Li, E. (1999). DNA Methyltransferases Dnmt3a and Dnmt3b Are Essential for De Novo Methylation and Mammalian Development. Cell *99*, 247–257.

Oksuz, O., Narendra, V., Lee, C.-H., Descostes, N., LeRoy, G., Raviram, R., Blumenberg, L., Karch, K., Rocha, P.P., Garcia, B.A., et al. (2018). Capturing the Onset of PRC2-Mediated Repressive Domain Formation. Molecular Cell *70*, 1149-1162.e5.

Onoshima, D., Kawakita, N., Takeshita, D., Niioka, H., Yukawa, H., Miyake, J., and Baba, Y. (2017). Measurement of DNA Length Changes upon CpG Hypermethylation by Microfluidic Molecular Stretching. Cell Med *9*, 61–66.

Ou, H.D., Phan, S., Deerinck, T.J., Thor, A., Ellisman, M.H., and O'Shea, C.C. (2017). ChromEMT: Visualizing 3D chromatin structure and compaction in interphase and mitotic cells. Science *357*, eaag0025.

Paliou, C., Guckelberger, P., Schöpflin, R., Heinrich, V., Esposito, A., Chiariello, A.M., Bianco, S., Annunziatella, C., Helmuth, J., Haas, S., et al. (2019). Preformed chromatin topology assists transcriptional robustness of Shh during limb development. PNAS *116*, 12390–12399.

Pasini, D., Bracken, A.P., Jensen, M.R., Denchi, E.L., and Helin, K. (2004). Suz12 is essential for mouse development and for EZH2 histone methyltransferase activity. The EMBO Journal *23*, 4061–4071.

Pengelly, A.R., Kalb, R., Finkl, K., and Müller, J. (2015). Transcriptional repression by PRC1 in the absence of H2A monoubiquitylation. Genes Dev. *29*, 1487–1492.

Perino, M., Mierlo, G. van, Karemaker, I.D., Genesen, S. van, Vermeulen, M., Marks, H., Heeringen, S.J. van, and Veenstra, G.J.C. (2018). MTF2 recruits Polycomb Repressive Complex 2 by helical-shape-selective DNA binding. Nat Genet *50*, 1002–1010.

Plys, A.J., Davis, C.P., Kim, J., Rizki, G., Keenen, M.M., Marr, S.K., and Kingston, R.E. (2019). Phase separation of Polycomb-repressive complex 1 is governed by a charged disordered region of CBX2. Genes Dev.

Poepsel, S., Kasinath, V., and Nogales, E. (2018). Cryo-EM structures of PRC2 simultaneously engaged with two functionally distinct nucleosomes. Nat Struct Mol Biol *25*, 154–162.

Pradeepa, M.M., Grimes, G.R., Kumar, Y., Olley, G., Taylor, G.C.A., Schneider, R., and Bickmore, W.A. (2016). Histone H3 globular domain acetylation identifies a new class of enhancers. Nat Genet *48*, 681–686.

Pugacheva, E.M., Kubo, N., Loukinov, D., Tajmul, M., Kang, S., Kovalchuk, A.L., Strunnikov, A.V., Zentner, G.E., Ren, B., and Lobanenkov, V.V. (2020). CTCF mediates chromatin looping via N-terminal domain-dependent cohesin retention. PNAS.

Qi, X., Li, T.-G., Hao, J., Hu, J., Wang, J., Simmons, H., Miura, S., Mishina, Y., and Zhao, G.-Q. (2004). BMP4 supports self-renewal of embryonic stem cells by inhibiting mitogen-activated protein kinase pathways. PNAS *101*, 6027–6032.

Ramani, V., Deng, X., Qiu, R., Gunderson, K.L., Steemers, F.J., Disteche, C.M., Noble, W.S., Duan, Z., and Shendure, J. (2017). Massively multiplex single-cell Hi-C. Nat Meth *14*, 263–266.

Ramírez, F., Bhardwaj, V., Arrigoni, L., Lam, K.C., Grüning, B.A., Villaveces, J., Habermann, B., Akhtar, A., and Manke, T. (2018). High-resolution TADs reveal DNA sequences underlying genome organization in flies. Nature Communications *9*, 189.

Rao, S., Huang, S.-C., Hilaire, B.G.S., Engreitz, J.M., Perez, E.M., Kieffer-Kwon, K.-R., Sanborn, A.L., Johnstone, S.E., Bochkov, I.D., Huang, X., et al. (2017a). Cohesin Loss Eliminates All Loop Domains, Leading To Links Among Superenhancers And Downregulation Of Nearby Genes. BioRxiv 139782.

Rao, S.S.P., Huntley, M.H., Durand, N.C., Stamenova, E.K., Bochkov, I.D., Robinson, J.T., Sanborn, A.L., Machol, I., Omer, A.D., Lander, E.S., et al. (2014). A 3D Map of the Human Genome at Kilobase Resolution Reveals Principles of Chromatin Looping. Cell *159*, 1665–1680.

Rao, S.S.P., Huang, S.-C., Glenn St Hilaire, B., Engreitz, J.M., Perez, E.M., Kieffer-Kwon, K.-R., Sanborn, A.L., Johnstone, S.E., Bascom, G.D., Bochkov, I.D., et al. (2017b). Cohesin Loss Eliminates All Loop Domains. Cell *171*, 305-320.e24.

Rauch, T.A., Wu, X., Zhong, X., Riggs, A.D., and Pfeifer, G.P. (2009). A human B cell methylome at 100–base pair resolution. PNAS *106*, 671–678.

Razin, S.V., and Gavrilov, A.A. (2014). Chromatin without the 30-nm fiber: Constrained disorder instead of hierarchical folding. Epigenetics *9*.

Reddington, J.P., Perricone, S.M., Nestor, C.E., Reichmann, J., Youngson, N.A., Suzuki, M., Reinhardt, D., Dunican, D.S., Prendergast, J.G., Mjoseng, H., et al. (2013). Redistribution of H3K27me3 upon DNA hypomethylation results in de-repression of Polycomb target genes. Genome Biology *14*, R25.

Reddington, J.P., Sproul, D., and Meehan, R.R. (2014). DNA methylation reprogramming in cancer: Does it act by re-configuring the binding landscape of Polycomb repressive complexes? BioEssays *36*, 134–140.

Reik, W. (2007). Stability and flexibility of epigenetic gene regulation in mammalian development. Nature *447*, 425–432.

Reverón-Gómez, N., González-Aguilera, C., Stewart-Morgan, K.R., Petryk, N., Flury, V., Graziano, S., Johansen, J.V., Jakobsen, J.S., Alabert, C., and Groth, A. (2018). Accurate Recycling of Parental Histones Reproduces the Histone Modification Landscape during DNA Replication. Mol. Cell *72*, 239-249.e5.

Rhodes, J.D.P., Feldmann, A., Hernández-Rodríguez, B., Díaz, N., Brown, J.M., Fursova, N.A., Blackledge, N.P., Prathapan, P., Dobrinic, P., Huseyin, M.K., et al. (2020). Cohesin Disrupts Polycomb-Dependent Chromosome Interactions in Embryonic Stem Cells. Cell Reports *30*, 820-835.e10.

Riggs, A.D. (1975). X inactivation, differentiation, and DNA methylation. CGR *14*, 9–25.

Riising, E.M., Comet, I., Leblanc, B., Wu, X., Johansen, J.V., and Helin, K. (2014). Gene Silencing Triggers Polycomb Repressive Complex 2 Recruitment to CpG Islands Genome Wide. Molecular Cell *55*, 347–360.

Robertson, E., Bradley, A., Kuehn, M., and Evans, M. (1986). Germ-line transmission of genes introduced into cultured pluripotential cells by retroviral vector. Nature *323*, 445–448.

Rose, N.R., King, H.W., Blackledge, N.P., Fursova, N.A., Ember, K.J., Fischer, R., Kessler, B.M., and Klose, R.J. (2016). RYBP stimulates PRC1 to shape chromatin-based communication between Polycomb repressive complexes. ELife *5*, e18591.

Ross, S.E., and Bogdanovic, O. (2019). TET enzymes, DNA demethylation and pluripotency. Biochemical Society Transactions BST20180606.

Rothbart, S.B., Krajewski, K., Nady, N., Tempel, W., Xue, S., Badeaux, A.I., Barsyte-Lovejoy, D., Martinez, J.Y., Bedford, M.T., Fuchs, S.M., et al. (2012). Association of UHRF1 with methylated H3K9 directs the maintenance of DNA methylation. Nature Structural & Molecular Biology *19*, 1155–1160.

Rowley, M.J., Lyu, X., Rana, V., Ando-Kuri, M., Karns, R., Bosco, G., and Corces, V.G. (2019). Condensin II Counteracts Cohesin and RNA Polymerase II in the Establishment of 3D Chromatin Organization. Cell Reports *26*, 2890-2903.e3.

Sanborn, A.L., Rao, S.S.P., Huang, S.-C., Durand, N.C., Huntley, M.H., Jewett, A.I., Bochkov, I.D., Chinnappan, D., Cutkosky, A., Li, J., et al. (2015). Chromatin extrusion explains key features of loop and domain formation in wild-type and engineered genomes. Proc Natl Acad Sci U S A *112*, E6456–E6465.

Santos, F., Peat, J., Burgess, H., Rada, C., Reik, W., and Dean, W. (2013). Active demethylation in mouse zygotes involves cytosine deamination and base excision repair. Epigenetics & Chromatin *6*, 39.

Sato, N., Meijer, L., Skaltsounis, L., Greengard, P., and Brivanlou, A.H. (2004). Maintenance of pluripotency in human and mouse embryonic stem cells through activation of Wnt signaling by a pharmacological GSK-3-specific inhibitor. Nat Med *10*, 55–63.

Saurin, A.J., Shiels, C., Williamson, J., Satijn, D.P.E., Otte, A.P., Sheer, D., and Freemont, P.S. (1998). The Human Polycomb Group Complex Associates with Pericentromeric Heterochromatin to Form a Novel Nuclear Domain. The Journal of Cell Biology *142*, 887–898.

Saurin, A.J., Shao, Z., Erdjument-Bromage, H., Tempst, P., and Kingston, R.E. (2001). A Drosophila Polycomb group complex includes Zeste and dTAFII proteins. Nature *412*, 655–660.

Sawyer, I.A., Bartek, J., and Dundr, M. (2019). Phase separated microenvironments inside the cell nucleus are linked to disease and regulate epigenetic state, transcription and RNA processing. Seminars in Cell & Developmental Biology *90*, 94–103.

Schmitges, F.W., Prusty, A.B., Faty, M., Stützer, A., Lingaraju, G.M., Aiwazian, J., Sack, R., Hess, D., Li, L., Zhou, S., et al. (2011). Histone Methylation by PRC2 Is Inhibited by Active Chromatin Marks. Molecular Cell *42*, 330–341.

Schoeftner, S., Sengupta, A.K., Kubicek, S., Mechtler, K., Spahn, L., Koseki, H., Jenuwein, T., and Wutz, A. (2006). Recruitment of PRC1 function at the initiation of X inactivation independent of PRC2 and silencing. The EMBO Journal *25*, 3110–3122.

Schoenfelder, S., Sugar, R., Dimond, A., Javierre, B.-M., Armstrong, H., Mifsud, B., Dimitrova, E., Matheson, L., Tavares-Cadete, F., Furlan-Magaril, M., et al. (2015). Polycomb repressive complex PRC1 spatially constrains the mouse embryonic stem cell genome. Nature Genetics *47*, 1179–1186.

Schübeler, D. (2015). Function and information content of DNA methylation. Nature *517*, 321–326.

Schuettengruber, B., Bourbon, H.-M., Di Croce, L., and Cavalli, G. (2017). Genome Regulation by Polycomb and Trithorax: 70 Years and Counting. Cell *171*, 34–57.

Schwarzer, W., Abdennur, N., Goloborodko, A., Pekowska, A., Fudenberg, G., Loe-Mie, Y., Fonseca, N.A., Huber, W., H. Haering, C., Mirny, L., et al. (2017). Two independent modes of chromatin organization revealed by cohesin removal. Nature *551*, 51–56.

Seisenberger, S., Andrews, S., Krueger, F., Arand, J., Walter, J., Santos, F., Popp, C., Thienpont, B., Dean, W., and Reik, W. (2012). The Dynamics of Genome-wide DNA Methylation Reprogramming in Mouse Primordial Germ Cells. Molecular Cell *48*, 849–862.

Servant, N., Varoquaux, N., Lajoie, B.R., Viara, E., Chen, C.-J., Vert, J.-P., Heard, E., Dekker, J., and Barillot, E. (2015). HiC-Pro: an optimized and flexible pipeline for Hi-C data processing. Genome Biology *16*, 259.

Shao, Z., Raible, F., Mollaaghababa, R., Guyon, J.R., Wu, C., Bender, W., and Kingston, R.E. (1999). Stabilization of Chromatin Structure by PRC1, a Polycomb Complex. Cell *98*, 37–46.

Shen, J.-C., Rideout, W.M., and Jones, P.A. (1994). The rate of hydrolytic deamination of 5-methylcytosine in double-stranded DNA. Nucleic Acids Res *22*, 972–976.

Shizuya, H., Birren, B., Kim, U.J., Mancino, V., Slepak, T., Tachiiri, Y., and Simon, M. (1992). Cloning and stable maintenance of 300-kilobase-pair fragments of human DNA in Escherichia coli using an F-factor-based vector. Proc Natl Acad Sci U S A *89*, 8794–8797.

Shopland, L.S., Lynch, C.R., Peterson, K.A., Thornton, K., Kepper, N., Hase, J. von, Stein, S., Vincent, S., Molloy, K.R., Kreth, G., et al. (2006). Folding and organization of a contiguous chromosome region according to the gene distribution pattern in primary genomic sequence. The Journal of Cell Biology *174*, 27–38.

Silva, J., Barrandon, O., Nichols, J., Kawaguchi, J., Theunissen, T.W., and Smith, A. (2008). Promotion of Reprogramming to Ground State Pluripotency by Signal Inhibition. PLOS Biology *6*, e253.

Simonis, M., Klous, P., Splinter, E., Moshkin, Y., Willemsen, R., de Wit, E., van Steensel, B., and de Laat, W. (2006). Nuclear organization of active and inactive chromatin domains uncovered by chromosome conformation capture–on-chip (4C). Nature Genetics *38*, 1348–1354.

Skene, P.J., and Henikoff, S. (2017). An efficient targeted nuclease strategy for high-resolution mapping of DNA binding sites. ELife *6*, e21856.

Smeets, D., Markaki, Y., Schmid, V.J., Kraus, F., Tattermusch, A., Cerase, A., Sterr, M., Fiedler, S., Demmerle, J., Popken, J., et al. (2014). Three-dimensional super-resolution microscopy of the inactive X chromosome territory reveals a collapse of its active nuclear compartment harboring distinct Xist RNA foci. Epigenetics & Chromatin *7*, 8.

Smith, T.A., and Hooper, M.L. (1983). Medium conditioned by feeder cells inhibits the differentiation of embryonal carcinoma cultures. Experimental Cell Research *145*, 458–462.

Smith, A.G., Heath, J.K., Donaldson, D.D., Wong, G.G., Moreau, J., Stahl, M., and Rogers, D. (1988). Inhibition of pluripotential embryonic stem cell differentiation by purified polypeptides. Nature *336*, 688–690.

Son, J., Shen, S.S., Margueron, R., and Reinberg, D. (2013). Nucleosome-binding activities within JARID2 and EZH1 regulate the function of PRC2 on chromatin. Genes Dev. *27*, 2663–2677.

Song, F., Chen, P., Sun, D., Wang, M., Dong, L., Liang, D., Xu, R.-M., Zhu, P., and Li, G. (2014). Cryo-EM Study of the Chromatin Fiber Reveals a Double Helix Twisted by Tetranucleosomal Units. Science *344*, 376–380.

Stevens, T.J., Lando, D., Basu, S., Atkinson, L.P., Cao, Y., Lee, S.F., Leeb, M., Wohlfahrt, K.J., Boucher, W., O'Shaughnessy-Kirwan, A., et al. (2017). 3D structures of individual mammalian genomes studied by single-cell Hi-C. Nature *544*, 59–64.

Strom, A.R., Emelyanov, A.V., Mir, M., Fyodorov, D.V., Darzacq, X., and Karpen, G.H. (2017). Phase separation drives heterochromatin domain formation. Nature *547*, 241–245.

Suetake, I., Shinozaki, F., Miyagawa, J., Takeshima, H., and Tajima, S. (2004). DNMT3L Stimulates the DNA Methylation Activity of Dnmt3a and Dnmt3b through a Direct Interaction. J. Biol. Chem. *279*, 27816–27823.

Tai, C.-I., and Ying, Q.-L. (2013). Gbx2, a LIF/Stat3 target, promotes reprogramming to and retention of the pluripotent ground state. J Cell Sci *126*, 1093–1098.

Tan, L., Xing, D., Chang, C.-H., Li, H., and Xie, X.S. (2018). Three-dimensional genome structures of single diploid human cells. Science *361*, 924–928.

Tatavosian, R., Kent, S., Brown, K., Yao, T., Duc, H.N., Huynh, T.N., Zhen, C.Y., Ma, B., Wang, H., and Ren, X. (2018). Nuclear condensates of the Polycomb protein chromobox 2 (CBX2) assemble through phase separation. J. Biol. Chem. jbc.RA118.006620.

Tavares, L., Dimitrova, E., Oxley, D., Webster, J., Poot, R., Demmers, J., Bezstarosti, K., Taylor, S., Ura, H., Koide, H., et al. (2012). RYBP-PRC1 Complexes Mediate H2A Ubiquitylation at Polycomb Target Sites Independently of PRC2 and H3K27me3. Cell *148*, 664–678.

Therizols, P., Illingworth, R.S., Courilleau, C., Boyle, S., Wood, A.J., and Bickmore, W.A. (2014). Chromatin decondensation is sufficient to alter nuclear organization in embryonic stem cells. Science *346*, 1238–1242.

Thompson, S., Clarke, A.R., Pow, A.M., Hooper, M.L., and Melton, D.W. (1989). Germ line transmission and expression of a corrected HPRT gene produced by gene targeting in embryonic stem cells. Cell *56*, 313–321.

Thomson, J.P., Skene, P.J., Selfridge, J., Clouaire, T., Guy, J., Webb, S., Kerr, A.R.W., Deaton, A., Andrews, R., James, K.D., et al. (2010). CpG islands influence chromatin structure via the CpG-binding protein Cfp1. Nature *464*, 1082–1086.

Tropberger, P., Pott, S., Keller, C., Kamieniarz-Gdula, K., Caron, M., Richter, F., Li, G., Mittler, G., Liu, E.T., Bühler, M., et al. (2013). Regulation of Transcription through Acetylation of H3K122 on the Lateral Surface of the Histone Octamer. Cell *152*, 859–872.

Tsukada, Y., Fang, J., Erdjument-Bromage, H., Warren, M.E., Borchers, C.H., Tempst, P., and Zhang, Y. (2006). Histone demethylation by a family of JmjC domain-containing proteins. Nature *439*, 811–816.

Ulianov, S.V., Tachibana-Konwalski, K., and Razin, S.V. (2017). Single-cell Hi-C bridges microscopy and genome-wide sequencing approaches to study 3D chromatin organization. BioEssays *39*, 1700104.

Vardimon, L., Kressmann, A., Cedar, H., Maechler, M., and Doerfler, W. (1982). Expression of a cloned adenovirus gene is inhibited by in vitro methylation. PNAS *79*, 1073–1077.

Vieux-Rochas, M., Fabre, P.J., Leleu, M., Duboule, D., and Noordermeer, D. (2015). Clustering of mammalian Hox genes with other H3K27me3 targets within an active nuclear domain. PNAS *112*, 4672–4677.

von Meyenn, F., Iurlaro, M., Habibi, E., Liu, N.Q., Salehzadeh-Yazdi, A., Santos, F., Petrini, E., Milagre, I., Yu, M., Xie, Z., et al. (2016). Impairment of DNA Methylation Maintenance Is the Main Cause of Global Demethylation in Naive Embryonic Stem Cells. Molecular Cell *62*, 848–861.

Walt, S. van der, Colbert, S.C., and Varoquaux, G. (2011). The NumPy Array: A Structure for Efficient Numerical Computation. Computing in Science Engineering *13*, 22–30.

Wang, H., Wang, L., Erdjument-Bromage, H., Vidal, M., Tempst, P., Jones, R.S., and Zhang, Y. (2004a). Role of histone H2A ubiquitination in Polycomb silencing. Nature *431*, 873–878.

Wang, L., Brown, J.L., Cao, R., Zhang, Y., Kassis, J.A., and Jones, R.S. (2004b). Hierarchical Recruitment of Polycomb Group Silencing Complexes. Molecular Cell *14*, 637–646.

Wang, R., Taylor, A.B., Leal, B.Z., Chadwell, L.V., Ilangovan, U., Robinson, A.K., Schirf, V., Hart, P.J., Lafer, E.M., Demeler, B., et al. (2010). Polycomb Group Targeting through Different Binding Partners of RING1B C-Terminal Domain. Structure *18*, 966–975.

Wang, S., Su, J.-H., Beliveau, B.J., Bintu, B., Moffitt, J.R., Wu, C., and Zhuang, X. (2016). Spatial organization of chromatin domains and compartments in single chromosomes. Science *353*, 598–602.

Wang, X., Paucek, R.D., Gooding, A.R., Brown, Z.Z., Ge, E.J., Muir, T.W., and Cech, T.R. (2017a). Molecular analysis of PRC2 recruitment to DNA in chromatin and its inhibition by RNA. Nature Structural & Molecular Biology *24*, 1028–1038.

Wang, X.-T., Cui, W., and Peng, C. (2017b). HiTAD: detecting the structural and functional hierarchies of topologically associating domains from chromatin interactions. Nucleic Acids Res *45*, e163–e163.

Wani, A.H., Boettiger, A.N., Schorderet, P., Ergun, A., Münger, C., Sadreyev, R.I., Zhuang, X., Kingston, R.E., and Francis, N.J. (2016). Chromatin topology is coupled to Polycomb group protein subnuclear organization. Nat Commun *7*.

Waters, T.R., and Swann, P.F. (2000). Thymine-DNA glycosylase and G to A transition mutations at CpG sites. Mutation Research/Reviews in Mutation Research *462*, 137–147.

Weide, R. van der (2019). GENome Organisation Visual Analytics.

Weinreb, C., and Raphael, B.J. (2016). Identification of hierarchical chromatin domains. Bioinformatics *32*, 1601–1609.

Wilder, P.J., Kelly, D., Brigman, K., Peterson, C.L., Nowling, T., Gao, Q.-S., McComb, R.D., Capecchi, M.R., and Rizzino, A. (1997). Inactivation of the FGF-4 Gene in Embryonic Stem Cells Alters the Growth and/or the Survival of Their Early Differentiated Progeny. Developmental Biology *192*, 614–629.

Wiles, M.V., and Johansson, B.M. (1999). Embryonic Stem Cell Development in a Chemically Defined Medium. Experimental Cell Research *247*, 241–248.

Williamson, I., Berlivet, S., Eskeland, R., Boyle, S., Illingworth, R.S., Paquette, D., Dostie, J., and Bickmore, W.A. (2014). Spatial genome organization: contrasting views from chromosome conformation capture and fluorescence in situ hybridization. Genes & Development *28*, 2778–2791.

Williamson, I., Lettice, L.A., Hill, R.E., and Bickmore, W.A. (2016). Shh and ZRS enhancer colocalisation is specific to the zone of polarising activity. Development *143*, 2994–3001.

Williamson, I., Kane, L., Devenney, P.S., Flyamer, I.M., Anderson, E., Kilanowski, F., Hill, R.E., Bickmore, W.A., and Lettice, L.A. (2019). Developmentally regulated Shh expression is robust to TAD perturbations. Development *146*, dev179523.

de Wit, E., Bouwman, B.A.M., Zhu, Y., Klous, P., Splinter, E., Verstegen, M.J.A.M., Krijger, P.H.L., Festuccia, N., Nora, E.P., Welling, M., et al. (2013). The pluripotent genome in three dimensions is shaped around pluripotency factors. Nature *501*, 227–231.

Woodcock, C.L., Skoultchi, A.I., and Fan, Y. (2006). Role of linker histone in chromatin structure and function: H1 stoichiometry and nucleosome repeat length. Chromosome Res. *14*, 17–25.

Wray, J., Kalkan, T., and Smith, A.G. (2010). The ground state of pluripotency. Biochemical Society Transactions *38*, 1027–1032.

Wu, H., Zeng, H., Dong, A., Li, F., He, H., Senisterra, G., Seitova, A., Duan, S., Brown, P.J., Vedadi, M., et al. (2013a). Structure of the Catalytic Domain of EZH2 Reveals Conformational Plasticity in Cofactor and Substrate Binding Sites and Explains Oncogenic Mutations. PLOS ONE *8*, e83737.

Wu, X., Johansen, J.V., and Helin, K. (2013b). Fbxl10/Kdm2b Recruits Polycomb Repressive Complex 1 to CpG Islands and Regulates H2A Ubiquitylation. Molecular Cell *49*, 1134–1146.

Yaffe, E., and Tanay, A. (2011). Probabilistic modeling of Hi-C contact maps eliminates systematic biases to characterize global chromosomal architecture. Nature Genetics *43*, 1059–1065.

Yamada, N.A., Rector, L.S., Tsang, P., Carr, E., Scheffer, A., Sederberg, M.C., Aston, M.E., Ach, R.A., Tsalenko, A., Sampas, N., et al. (2011). Visualization of Fine-Scale Genomic Structure by Oligonucleotide-Based High-Resolution FISH. CGR *132*, 248–254.

Yamaji, M., Ueda, J., Hayashi, K., Ohta, H., Yabuta, Y., Kurimoto, K., Nakato, R., Yamada, Y., Shirahige, K., and Saitou, M. (2013). PRDM14 Ensures Naive Pluripotency through Dual Regulation of Signaling and Epigenetic Pathways in Mouse Embryonic Stem Cells. Cell Stem Cell *12*, 368–382.

Yan, J., Dutta, B., Hee, Y.T., and Chng, W.-J. (2019). Towards understanding of PRC2 binding to RNA. RNA Biology *16*, 176–184.

Yin, Y., Morgunova, E., Jolma, A., Kaasinen, E., Sahu, B., Khund-Sayeed, S., Das, P.K., Kivioja, T., Dave, K., Zhong, F., et al. (2017). Impact of cytosine methylation on DNA binding specificities of human transcription factors. Science *356*, eaaj2239.

Ying, Q.-L., Nichols, J., Chambers, I., and Smith, A. (2003). BMP Induction of Id Proteins Suppresses Differentiation and Sustains Embryonic Stem Cell Self-Renewal in Collaboration with STAT3. Cell *115*, 281–292.

Ying, Q.-L., Wray, J., Nichols, J., Batlle-Morera, L., Doble, B., Woodgett, J., Cohen, P., and Smith, A. (2008). The ground state of embryonic stem cell self-renewal. Nature *453*, 519–523.

Yu, J.-R., Lee, C.-H., Oksuz, O., Stafford, J.M., and Reinberg, D. (2019). PRC2 is high maintenance. Genes Dev. *33*, 903–935.

Zhang, Y., Xiang, Y., Yin, Q., Du, Z., Peng, X., Wang, Q., Fidalgo, M., Xia, W., Li, Y., Zhao, Z., et al. (2018). Dynamic epigenomic landscapes during early lineage specification in mouse embryos. Nature Genetics *50*, 96.

Zhao, Z., Tavoosidana, G., Sjölinder, M., Göndör, A., Mariano, P., Wang, S., Kanduri, C., Lezcano, M., Singh Sandhu, K., Singh, U., et al. (2006). Circular chromosome conformation capture (4C) uncovers extensive networks of epigenetically regulated intra- and interchromosomal interactions. Nature Genetics *38*, 1341–1347.

Zhen, C.Y., Duc, H.N., Kokotovic, M., Phiel, C.J., and Ren, X. (2014). Cbx2 stably associates with mitotic chromosomes via a PRC2- or PRC1-independent mechanism and is needed for recruiting PRC1 complex to mitotic chromosomes. Mol Biol Cell *25*, 3726–3739.

Zhu, J., He, F., Hu, S., and Yu, J. (2008). On the nature of human housekeeping genes. Trends in Genetics *24*, 481–484.

Zufferey, M., Tavernari, D., Oricchio, E., and Ciriello, G. (2018). Comparison of computational methods for the identification of topologically associating domains. Genome Biology *19*, 217.

Zuo, Z., Roy, B., Chang, Y.K., Granas, D., and Stormo, G.D. (2017). Measuring quantitative effects of methylation on transcription factor–DNA binding affinity. Science Advances *3*, eaao1799.